ON PREDICTIVE INFERENCE

by

David Hinkley

University of Minnesota

Technical Report No. 254

January 1976

ON PREDICTIVE INFERENCE

by

David Hinkley
University of Minnesota

The likelihood function is the common basis of all parametric inference. However, with the exception of an ad hoc definition by R.A. Fisher, there is no such unifying basis for prediction of future events, given past observations. This article proposes a definition of predictive likelihood which can help to remove some non-uniqueness problems in sampling-theory predictive inference, and which can produce a simple prediction analog of the Bayes result, posterior $\propto$ prior x likelihood, in many situations.

AMS 1970 subject classifications.

Key words and phrases. Prediction, confidence regions, exponential families, likelihood, Bayesian methods, pivotal, cross-validation

1. Introduction

In 1920 Karl Pearson posed "The Fundamental Problem of Applied Statistics" as follows:

'An "event" has occurred p times out p + q = n trials, where we have no a priori knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further r + s = m trials?' Pearson's purpose was to reexamine the general applicability of Bayes's earlier solution, and the resulting controversy, described by Edwards [4], is of some interest. However, the main question seems to have been largely ignored in the intervening years, while parametric inference has occupied statistical thought. Some authors, for example de Finetti [7] and Geisser [11], have suggested that more attention be paid to prediction of variables.

Pearson's question could be interpreted in at least two ways. For example, if θ is the frequency of the "event" in the total population, so that the unknowable answer to the Fundamental Problem is $\binom{m}{r} \theta^r (1-\theta)^{m-r}$, then the question could be answered by giving a point estimate for this real frequency probability. This means interpreting the problem as a decision problem. Alternatively, we might view the problem as an inferential one, as Pearson and Fisher ([16],[17],[18]) did, for which the answer is in terms of relative credibilities for various values of r. Of course a Bayesian solution does this in terms of a different version of probability, and one might expect that other types of credibility exist for non-Bayesians.

It is to this inferential aspect of prediction that the present paper is directed. The main purpose is to present and discuss a definition

Introduction

In 1970 Neil Postman and [illegible] published a critique of television entitled "Teaching as a Subversive Activity" based on the premise that the school system was failing to prepare students adequately for the world they were going to enter. Postman and [illegible] argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

Postman and [illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

[illegible] (1971) also published a book entitled "Teaching as a Subversive Activity" in which they argued that the curriculum of the school system was based on the assumption that the world would remain relatively stable, whereas the world was actually changing rapidly. They suggested that the curriculum should be changed to reflect the changing world.

of predictive likelihood, analogous to parametric likelihood, which is Fisherian in spirit but which, strangely, was not suggested by Fisher. The two desirable properties of likelihood, namely resolution of non-unique frequentist confidence regions and ability to yield Bayesian methods, have motivated the definition. Quite by chance, S. Lauritzen and the author proposed much the same likelihood in early 1974, but our work has pursued different paths; Lauritzen's mathematical account has since appeared in [14]. The reader will find the present account deliberately non-mathematical in character; hopefully this will allow easy digestion of the statistical ideas.

Section 2 briefly reviews the two main sampling-theory methods of predictive inference which lead to confidence statements about future random variables as well as the Bayesian method of defining posterior predictive distributions. The review suggests the need for an analog of parametric likelihood. In Section 3 the analog is given, and shown to be a factor in the Bayes posterior predictive distribution. There is also some discussion of the likelihood as a credibility measure.

of predictive likelihood, analogous to parametric likelihood, which is Fisherian in spirit but, which, strangely, was not suggested by Fisher. The two desirable properties of likelihood, namely resolution of non-unique frequentist confidence regions and ability to yield Bayesian methods, have motivated the definition. Quite by chance, S. Lauritzen and the author proposed much the same likelihood in early 1974, but our work has pursued different paths; Lauritzen's mathematical account has since appeared in (Lauritzen, 1974). The reader will find the present account deliberately non-mathematical in outlook; hopefully this will allow easy digestion of the statistical ideas.

Section 2 briefly reviews the two main sampling-theory methods of predictive inference which lead to confidence statements about future random variables, as well as the Bayesian method of obtaining posterior predictive distributions. The review suggests the need for an analog of parametric likelihood. In Section 3 the analog is given, and shown to be a factor in the Bayes posterior predictive distribution. There is also some discussion of the likelihood as a credibility measure.

2. A Brief Critique of Standard Forms of Predictive Inference

2.1 Introduction

Suppose that $Y = (Y_1,\ldots,Y_n)$ is a vector random variable with joint probability density function, p.d.f., $f_Y(y;\theta)$, $\theta \in \Omega$, and that the random variable $Z = (Z_1,\ldots,Z_m)$ independently has p.d.f. $g_Z(z;\theta)$. An observation y on Y is available, and we wish to predict Z in the absence of knowledge about θ. The independence of Y and Z is assumed only for simplicity, and the general theory outlined below extends easily.

Prediction of Z could mean many things. It could mean estimation of the p.d.f. g_Z, or its probability integral. Or, prediction could mean construction of a series of confidence regions. In the absence of a utility specification, this latter interpretation is the usual one, which we discuss here. First, in Section 2.2, the two main sampling-theory types of confidence region (critical region and pivotal) are described, and their non-uniqueness emphasized. Section 2.3 defines the Bayesian posterior predictive distribution and suggests the desirability of a prediction analog of parametric likelihood.

2.2 Sampling-theory Predictive Inference

A confidence region for Z is a set $P_\alpha(Y)$ in the sample space of Z, determined by the observable Y and satisfying

$$\text{pr}\{Z \in P_\alpha(Y);\ \theta\} = 1 - \alpha \qquad (2.1)$$

independently of θ; several regions, for different values of α, might be given simultaneously. There are two main methods of constructing $P_\alpha(Y)$, one based on test critical regions, the other based on pivotal quantities.

Construction of $P_\alpha(Y)$ via test critical regions proceeds as follows: Suppose that Z has p.d.f. $g_Z(z;\theta^*)$, and consider the hypothesis $H_o:\theta = \theta^*$

1.5 Modèles

Soient deux échantillons aléatoires notés $(X_1,\ldots,X_m) = X$ et $(Y_1,\ldots,Y_n) = Y$ issus respectivement des lois $F(\cdot)$ et $G(\cdot)$, c.à.d.q ayant fonctions de répartition absolument continues. Nous supposerons de plus que $F(x) \geq G(x)$ $\forall x \in \mathbb{R}$, ce qui signifie que les valeurs prises par X sont en général plus grandes que celles prises par Y. Nous nous proposons de tester l'hypothèse nulle $H_0 : F = G$ contre l'alternative $H_1 : F \neq G$ ou $H_1 : F > G$ ou $H_1 : F < G$ selon le cas.

Nous nous proposons de tester H_0 contre H_1 en utilisant la statistique $D_{m,n}$ définie par $D_{m,n} = \sup_x |F_m(x) - G_n(x)|$ où F_m et G_n sont les fonctions de répartition empiriques associées à F et G respectivement. Nous rejetterons H_0 au profit de H_1 si $D_{m,n}$ est trop grande.

Nous nous proposons de tester H_0 contre H_1 en utilisant la statistique $D_{m,n}$ définie par $D_{m,n} = \sup_x |F_m(x) - G_n(x)|$ où F_m et G_n sont les fonctions de répartition empiriques associées à F et G respectivement. Nous rejetterons H_0 au profit de H_1 si $D_{m,n}$ est trop grande.

1.6 Quelques résultats classiques sur la statistique $D_{m,n}$

Nous nous proposons de tester H_0 contre H_1 en utilisant la statistique $D_{m,n}$ définie par $D_{m,n} = \sup_x |F_m(x) - G_n(x)|$ où F_m et G_n sont les fonctions de répartition empiriques associées à F et G respectivement. Nous rejetterons H_0 au profit de H_1 si $D_{m,n}$ est trop grande.

Nous nous proposons de tester H_0 contre H_1 en utilisant la statistique $D_{m,n}$ définie par $D_{m,n} = \sup_x |F_m(x) - G_n(x)|$ où F_m et G_n sont les fonctions de répartition empiriques associées à F et G respectivement. Nous rejetterons H_0 au profit de H_1 si $D_{m,n}$ est trop grande.

with some alternative hypothesis H_A. If for the particular H_A a size - α similar test critical region can be found, then by definition

$$\text{pr}\{(Y, Z) \in R_\alpha ; H_o\} = \alpha . \tag{2.2}$$

But H_o <u>is</u> true, and so if $P_\alpha(y)$ is the projection of the complement $\overline{R}_\alpha$ onto $Y = y$, then $P_\alpha(Y)$ satisfies the requirement (2.1). Thus the prediction confidence region contains all values of z which are not in conflict with the model and observation y, as judged by a size - α test.

It is clear that many such R_α may exist, and their existence is not dependent on the nature of H_A. However, the standard procedure is to determine a "good" critical region for the particular alternative H_A introduced, in the sense of being most powerful (locally or uniformly). For many problems this results in precise confidence statements about Z, but the general concept of power seems somewhat remote from prediction.

Two important points to note about this approach are:

(i) If the minimal sufficient statistic under H_o is not a reduction of the minimal sufficient statistics for the two families, then the Neyman structure property of similar tests is lost, and there may be no similar critical regions R_α. A simple example is that where the observables $Y_1,\ldots,Y_n$ are independent counts from a Poisson process of rate θ, and Z is the time from "now" to the occurrence of the next event.

(ii) The method is not uniquely defined in the sense that different alternatives H_A are possible. Indeed if $\dim(\theta) > \dim(Z)$, then there must be a restriction on H_A in order to obtain non-trivial critical regions, and there is no unique restriction. For example, if $Y_1,\ldots,Y_n$, are i.i.d.

with some alternative hypothesis H_A. If for the particular H_A a size-α similar test critical region can be found, then by definition

$$\mathrm{pr}\{(Y, Z) \in R \,;\, H_0\} = \alpha \,. \qquad (2.2)$$

But H_0 is true, and so if $Z_\alpha(y)$ is the projection of the complement $\bar{R}$ onto $Y = y$, then $Z_\alpha(Y)$ satisfies the requirement (2.1). Thus the prediction confidence region contains all values of z which are not in conflict with the model and observation y, as judged by a size-α test.

It is clear that many such R_α may exist, and their existence is not dependent on the nature of H_A. However, the standard procedure is to determine a "good" critical region for the particular alternative H introduced, in the sense of being most powerful (locally or uniformly). For many problems this results in precise confidence statements about Z, but the general concept of power seems somewhat remote from prediction.

Two important points to note about this approach are:

(i) If the minimal sufficient statistic under H_0 is not a reduction of the minimal sufficient statistics for the two families, then the Neyman structure property of similar tests is lost, and there may be no similar critical regions R_α. A simple example is that where the observations $Y_1,\ldots,Y_n$ are independent counts from a Poisson process of rate θ, and Z is the time from "now" to the occurrence of the next event.

(ii) The method is not uniquely defined in the sense that different alternatives H_A may be chosen. Indeed if $\dim(\phi)$ $\dim(\theta) > \dim(\zeta)$, then there must be a restriction on H_A in order to obtain non-trivial critical regions, and there is no unique restriction. For example, if $Y_1,\ldots,Y_n$ are i.i.d.

$N(\mu,\tau)$ and Z_1 is independently $N(\mu^*,\tau^*)$, then we cannot allow both $\mu \neq \mu^*$ and $\tau \neq \tau^*$; the form of the prediction confidence region will depend on H_A if a most powerful critical region is used.

The above points are made to imply a degree of arbitrariness in the critical region construct. Similar criticisms can be made against critical region construction of confidence regions for θ, but in that case non-uniqueness is removable by use of the likelihood function. That is to say, among the systems of confidence regions $C_\alpha(Y)$ for θ which exist with the required property

$$\text{pr}\{\theta \in C_\alpha(Y);\ \theta\} = 1 - \alpha ,$$

we take the uniquely-defined likelihood-based confidence region $C^*_\alpha(y)$ satisfying

$$\inf_{\theta' \in C^*_\alpha(y)} \text{lik}(\theta'|y) \geq \sup_{\theta'' \notin C^*_\alpha(y)} \text{lik}(\theta''|y) . \tag{2.3}$$

It seems a reasonable requirement that no parameter point outside the confidence region have higher likelihood than any point inside the region, and a corresponding likelihood base for our predictand Z is much to be desired.

The other main method for obtaining prediction confidence regions is the use of pivotal quantities. This has strong connections with fiducial and structural inference, but we shall not pursue them here. Briefly, a pivotal quantity in the prediction context is a function $r(Y, Z)$ with distribution independent of θ. Thus for any α, we can in principle determine a region T_α such that

$$\text{pr}\{r(Y, Z) \in T_\alpha;\ \theta\} = 1 - \alpha \tag{2.4}$$

$N(\mu, \sigma)$ and Z_1 is independently $N(\mu, \sigma^2)$; the form of the prediction confidence region will depend on H_1 (if a most powerful critical region is used).. if a most powerful critical region is used.

The above points are made to imply a degree of arbitrariness in the critical region construct. Similar criticisms can be made against critical region construction of confidence regions for θ, but in that case non-uniqueness is removable by use of the likelihood function. That is to say, among the several systems of confidence regions $C_\alpha(Y)$ for θ which exist with the required property

$$\Pr\{\theta \in C_\alpha(Y); \theta\} = 1 - \alpha$$

we take the uniquely-defined likelihood-based confidence region $C^*_\alpha(y)$ satisfying

$$\inf_{\theta' \in C^*_\alpha(y)} \mathrm{lik}(\theta'|y) \geq \sup_{\theta'' \notin C^*_\alpha(y)} \mathrm{lik}(\theta''|y) . \qquad (2.3)$$

It seems a reasonable requirement that no parameter point outside the confidence region have higher likelihood than any point inside the region, and a corresponding likelihood base for our predictand Z is much to be desired.

The other main method for obtaining prediction confidence regions is the use of pivotal quantities. This has strong connections with fiducial and structural inference, but we shall not pursue those here. Briefly, a pivotal quantity in the prediction context is a function r(Y, Z) with distribution independent of θ. Thus for any α, we can in principle determine a region T_α such that

$$\mathrm{pr}\{r(Y, Z) \in T_\alpha ; \theta\} = 1 - \alpha \qquad (2.4)$$

for all θ; cf. (2.2). Then a pivotal prediction region is the projection of T_α onto $Y = y$,

$$P_\alpha(y) = \{z | r(y,z) \in T_\alpha\} . \tag{2.5}$$

This approach requires construction of an appropriate pivotal quantity $r(Y, Z)$, which can be accomplished by an extension of G.A. Barnard's work on pivotal parametric inference; see [1].

We retain the assumption of independence of Y and Z for simplicity. Then suppose that the one-one transformation γ takes Y into

$$C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} c_1(Y,\theta) \\ c_2(Y) \end{pmatrix} ,$$

where C_1 and C_2 are pivots having distributions independent of θ; C_1 is termed <u>proper</u>, and C_2 is termed <u>ancillary</u>. Suppose that a corresponding transformation δ takes Z into

$$D = \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} = \begin{pmatrix} d_1(Z,\theta) \\ d_2(Z) \end{pmatrix} ,$$

where D_1 and D_2 are respectively proper and ancillary pivotals. Since D_2 has known distribution, the prediction problem focusses on D_1, to which the information about θ in the conditional distribution of C_1 given $C_2 = c_2$ relates. Now we make the definition of predictive pivot: if a one-one transformation of (C_1,D_1) exists such that

$$\begin{pmatrix} C_1 \\ D_1 \end{pmatrix} \to \begin{pmatrix} Q \\ R \end{pmatrix} = \begin{pmatrix} q(Y,Z,\theta) \\ r(Y,Z) \end{pmatrix}$$

then R is a <u>predictive pivot</u>. With the transformation of (Y,Z) into (C_2,D_2,Q,R) is associated the re-expression of their joint distribution,

notación es la misma que la usada en (4.3). Además hacemos $\xi = \nu$ y $\eta = \chi$ en (4.3) para obtener

$$\Pi_{\nu}^{\chi}(\chi) = [\pi/\pi(\chi)\,;\ \Pi_{\nu}^{\chi}] \qquad (4.4)$$

Esta expresión establece que la función característica $\Pi_{\nu}^{\chi}(\chi)$ es igual a la probabilidad de que χ pertenezca a Π_{ν}^{χ} dado que χ pertenece a $\pi(\chi)$. Además, (4.4) implica que χ pertenece a Π_{ν}^{χ} si y sólo si χ pertenece a $\pi(\chi)$. Esto es, $\Pi_{\nu}^{\chi}(\chi) = 1$ si y sólo si $\chi \in \pi(\chi)$.

Ahora bien, $\pi(\chi)$ es un conjunto de vectores de la forma $(\chi, \cdot)$. Por lo tanto, $\chi \in \pi(\chi)$ si y sólo si existe un vector $(\chi, \cdot)$ en $\pi(\chi)$. Esto es, $\chi \in \pi(\chi)$ si y sólo si χ es la primera componente de algún vector en $\pi(\chi)$. Por lo tanto, $\Pi_{\nu}^{\chi}(\chi) = 1$ si y sólo si χ es la primera componente de algún vector en $\pi(\chi)$. Esto es, $\Pi_{\nu}^{\chi}(\chi) = 1$ si y sólo si existe un vector $(\chi, \cdot)$ en $\pi(\chi)$.

Ahora bien, $\pi(\chi)$ es el conjunto de todos los vectores $(\chi, \cdot)$ que pertenecen a Π_{ν}^{χ}. Por lo tanto, $\Pi_{\nu}^{\chi}(\chi) = 1$ si y sólo si existe un vector $(\chi, \cdot)$ en Π_{ν}^{χ}. Esto es, $\Pi_{\nu}^{\chi}(\chi) = 1$ si y sólo si existe un vector $(\chi, \cdot)$ en Π_{ν}^{χ}.

$$\begin{pmatrix} c^{\nu} \\ c^{\chi} \end{pmatrix} = \begin{pmatrix} c^{\nu}(\chi) \\ c^{\chi}(\chi, \cdot) \end{pmatrix}$$

Ahora bien, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ.

$$\begin{pmatrix} c^{\nu} \\ c^{\chi} \end{pmatrix} = \begin{pmatrix} c^{\nu}(\chi) \\ c^{\chi}(\chi, \cdot) \end{pmatrix}$$

Ahora bien, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ.

$$\begin{pmatrix} c^{\nu} \\ c^{\chi} \end{pmatrix} = \begin{pmatrix} c^{\nu}(\chi) \\ c^{\chi}(\chi, \cdot) \end{pmatrix}$$

Ahora bien, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ.

$$\begin{pmatrix} c^{\nu} \\ c^{\chi} \end{pmatrix} = \begin{pmatrix} c^{\nu}(\chi) \\ c^{\chi}(\chi, \cdot) \end{pmatrix}$$

Ahora bien, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ.

$$\begin{pmatrix} c^{\nu} \\ c^{\chi} \end{pmatrix} = \begin{pmatrix} c^{\nu}(\chi) \\ c^{\chi}(\chi, \cdot) \end{pmatrix}$$

Ahora bien, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ. Esto es, c^{ν} es el vector de Π_{ν}^{χ} que corresponde a χ.

which takes the form

$$f_Y(y;\theta)g_Z(z;\theta) = \varphi(c_2)\ \psi\ (d_2)\ \xi\ (r|c_2,d_2)\ \eta\ (q|c_2,d_2,r;\theta).$$

The prediction problem is then resolved by discarding $q(Y,Z,\theta)$, either on the principle that θ and Z are confounded, or by appeal to lack of invariance of Q under isomorphic transformations of Y and Z which leave R invariant. With Q discarded, we are left with the following predictive information:

> $c_2(Z)$ has p.d.f. $\varphi(c_2)$, and conditional on $(C_2,D_2) = (c_2,d_2)$ the pivot $r(Y,Z)$ has p.d.f. $\xi(r|c_2,d_2)$, the observed value of Y being y.

Note that this leaves ambiguous the particular choice of region T_α, and hence $P_\alpha(y)$, in (2.4). For example, in a location parameter problem where the predictive pivot R may be chosen as $\bar{Y} - \bar{Z} = n^{-1}\Sigma Y_j - m^{-1}\Sigma Z_j$, $P_\alpha(y)$ is any invariant region of size $1 - \alpha$ with respect to the conditional distribution of R, shifted by amount $\bar{y}$. The ambiguity would be removed by defining $\xi(r|c_2,d_2)\ \varphi\ (c_2)$ as the likelihood function of z given y, to be used to uniquely define $P^*_\alpha(y)$ in a manner analogous to $C^*_\alpha(y)$ in (2.3). This definition of likelihood is unsatisfactory on two counts: Firstly, since the pivotal approach fails for most discrete distributions, we would be left with no definition of likelihood. Second, this definition of likelihood would not generally permit combination with prior information about Z to obtain Bayes posterior predictive distributions; in fact explicit prior information needs to be represented by the pivotal quantity ⊛, which together with C_1 and D_1 defines the conditional distribution of Z; see [1].

which takes the form

$$L_1(y;\theta_2(y)) = f(c_2)\, f(d_2)\, f(u\,|\,c_2,d_2)\, f(q\,|\,c_2,d_2;u;\theta).$$

The prediction problem is then resolved by discarding $q(X,Z,\theta)$, either on the principle that θ and Z are confounded: or by appeal to lack of invariance of q under isomorphic transformations of X and Z which leave R invariant. With θ discarded, we are left with the following predictive information:

> $c_2(Z)$ has p.d.f. $f(c_2)$, and conditional on $(C_2,D_2) = (c_2,d_2)$ the pivot $u(X,Z)$ has p.d.f. $f(u|c_2,d_2)$, the observed value of X being y.

Note that this leaves ambiguous the particular choice of region R_y, and hence $P_\alpha(y)$, in (2.4). For example, in a location parameter problem where the predictive pivot may be chosen as $\bar{X} - \bar{Z} = n^{-1}\Sigma X_i - m^{-1}\Sigma Z_j$, $P_\alpha(y)$ is any invariant region of size $1-\alpha$ with respect to the conditional distribution of u, shifted by amount $\bar{y}$. The ambiguity would be removed by defining $f(u|c_2,d_2)\,f(c_2)$ as the likelihood function of a given u, to be used to uniquely define $P^*_\alpha(y)$ in a manner analogous to $P^*_\alpha(y)$ in (2.3). This definition of likelihood is unsatisfactory on two counts: Firstly, since the pivotal approach fails for most discrete distributions, we would be left with no definition of likelihood. Second, this definition of likelihood would not generally permit combination with prior information about θ to obtain Bayes posterior predictive distributions; in fact explicit prior information needs to be represented by the pivotal quantity θ, which together with C_1 and D_1 defines the conditional distribution of θ; see Barnard (1975).

To sum up: The two most useful approaches to construction of similar prediction confidence regions do not in fact uniquely define such regions, nor do they give solutions to all problems. Partial resolution of these difficulties requires definition of a suitably general predictive likelihood.

2.3 Bayesian Predictive Inference

The preceding critique of sampling-theory predictive inference would usually fall on deaf Bayesian ears. However, a prediction analog of parametric likelihood might be of some interest within the Bayesian framework. Recall that for the general case, if $p_{\Theta}(\theta)$ is the prior density for Θ, then the Bayes posterior predictive distribution will have density

$$f_{Z|Y}(z|y) = \int f_{\Theta|Y}(\theta|y) f_{Z|Y,\theta}(z|y,\theta)d\theta$$

$$= \frac{\int f_{Y|\Theta}(y|\theta)p_{\Theta}(\theta)f_{Z|Y,\Theta}(z|y,\theta)d\theta}{\int f_{Y|\Theta}(y|\theta)p_{\Theta}(\theta)d\theta} \quad . \tag{2.5}$$

This appears to be somewhat complicated when compared to the "loose" form of the Bayes posterior for Θ, i.e.

$$f_{\Theta|Y}(\theta|y) \propto \text{lik}_Y(\theta|y)p_{\Theta}(\theta) \quad .$$

It would, then, be of some interest to obtain a prediction analog of the likelihood $\text{lik}_Y(\theta|y)$ for which (2.5) is expressible as

$$f_{Z|Y}(z|y) \propto \text{likelihood} \times p_Z(z) \quad , \tag{2.6}$$

where $P_Z(z)$ is the marginal prior density of Z. Indeed in many ways (2.6) would seem to be a preferred form of (2.5), since presumably subjective probability assessments concerning z would often be more tractable than corresponding assessments concerning θ.

...ggestive of the presence of a hidden variable

...ere is dependence between the variables and

...dependence is said to be conditional (2.2)

...ther words if the value of λ is known then the ... (2.3)

...ecome independent. Thus we have ...

$$P(a|\lambda) = P(a|\lambda)P(b|\lambda) \quad (2.4)$$

...e correlation between the ... (2.5)

$$P(a,b) = \int P(a|\lambda)P(b|\lambda)\rho(\lambda)\,d\lambda \quad (2.6)$$

...ecause we do not know the ... (2.7)

$$P(a,b) = \frac{\int P(a|\lambda)P(b|\lambda)\rho(\lambda)\,d\lambda}{\int \rho(\lambda)\,d\lambda} \equiv \int P(a|\lambda)P(b|\lambda)\rho(\lambda)\,d\lambda \quad (2.8)$$

...hich we have assumed to be normalized to unity. ...

$$P(a,b) = \int P(a|\lambda)P(b|\lambda)\rho(\lambda)\,d\lambda \quad (2.9)$$

...ocal hidden variable theories ... Bell's theorem ...

...ere are two ways to ...

...ne may ...

3. Predictive Likelihoods

3.1 Introduction

The previous section suggests a need for the prediction analog of parametric likelihood. One such analog was introduced, rather naturally, by R.A. Fisher [10], although his use was rather ad hoc. Section 3.2 describes Fisher's definition and shows it to be defective in terms of reasonable consistency requirements. A new definition of predictive likelihood is proposed in Section 3.3, which details various important properties of the likelihood for exponential families, including the satisfaction of the Bayes factorization (2.6). Section 3.4 briefly discusses a conditional version of predictive likelihood that is operational for location-scale families, for example.

3.2 Fisher's Predictive Likelihood

In his book Statistical Methods and Scientific Inference Fisher discussed the "Fundamental Problem" described in Section 1. He proposed a likelihood for the future sample based on the likelihood ratio measure of support for the true hypothesis of constant probability that is obtained from comparison of the observed sample and the future sample. Specifically, if the event has occurred y times out of n trials, where the chance of occurrence is θ on each trial, and if z is the number of occurrences in m further trials, then the predictive likelihood is the likelihood ratio statistic for testing that θ has the same value in both samples, namely

$$\mathrm{FLIK}(z|y) = \frac{\sup_{\theta} \binom{m+n}{y+z} \theta^{y+z}(1-\theta)^{m+n-(y+z)}}{\sup_{\varphi} \binom{n}{y} \varphi^{y}(1-\varphi)^{n-y} \sup_{\varphi} \binom{m}{z} \varphi^{z}(1-\varphi)^{m-z}}$$

$$= \frac{(y+z)^{y+z}(m+n-y-z)^{m+n-y-z} n^{n} m^{m}}{y^{y}(n-y)^{n-y} z^{z}(m-z)^{m-z}} \,. \qquad (3.1)$$

c. Двойственная функция

1.6 Умножение

Пусть f и g — две произвольные булевы функции от n переменных. Произведение $f \cdot g$ — это булева функция от n переменных, которая принимает значение 1 на тех и только на тех наборах переменных, на которых обе функции f и g принимают значение 1. Произведение $f \cdot g$ называют также *конъюнкцией* функций f и g. Конъюнкцию обозначают также $f \wedge g$ или $f \& g$.

Таким образом, по определению

$$(f \cdot g)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \cdot g(x_1, \ldots, x_n).$$

Произведение $f \cdot g$ можно также обозначать fg, опуская точку умножения.

1.7 Дизъюнкция

Пусть f и g — две произвольные булевы функции от n переменных. Сумма $f + g$ по модулю 2 — это булева функция от n переменных, которая принимает значение 1 на тех и только на тех наборах переменных, на которых значение функции f не совпадает со значением функции g. Сумму $f + g$ называют также *дизъюнкцией* функций f и g. Дизъюнкцию обозначают также $f \vee g$ или $f \cup g$.

Таким образом, по определению

$$(f + g)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) + g(x_1, \ldots, x_n) \pmod 2.$$

Сумму $f + g$ по модулю 2 называют также *симметрической разностью* функций f и g. Симметрическую разность обозначают также $f \triangle g$ или $f \setminus g$.

Дизъюнкцию $f + g$ можно также обозначать $f \vee g$, если $f \cdot g = 0$, т.е. если функции f и g не принимают одновременно значение 1 ни на одном наборе переменных. В этом случае дизъюнкцию называют также *логической суммой* функций f и g.

1.8 Импликация

Импликацией функций f и g называется булева функция $f \to g$, которая принимает значение 0 на тех и только на тех наборах переменных, на которых $f = 1$ и $g = 0$. Таким образом, по определению

$$(f \to g)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \cdot g(x_1, \ldots, x_n) + f(x_1, \ldots, x_n) + 1 \pmod 2.$$

Импликацию $f \to g$ можно также обозначать $f \subset g$.

1.9 Эквивалентность

Эквивалентностью функций f и g называется булева функция $f \sim g$, которая принимает значение 1 на тех и только на тех наборах переменных, на которых значения функций f и g совпадают. Таким образом, по определению

$$(f \sim g)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) + g(x_1, \ldots, x_n) + 1 \pmod 2.$$

Эквивалентность $f \sim g$ можно также обозначать $f \equiv g$.

1.10 Штрих Шеффера

Штрихом Шеффера функций f и g называется булева функция $f \mid g$, которая принимает значение 0 на тех и только на тех наборах переменных, на которых $f = 1$ и $g = 1$. Таким образом, по определению

$$(f \mid g)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \cdot g(x_1, \ldots, x_n) + 1 \pmod 2.$$

Штрих Шеффера называют также *антиконъюнкцией* функций f и g.

1.11 Стрелка Пирса

Стрелкой Пирса функций f и g называется булева функция $f \downarrow g$, которая принимает значение 1 на тех и только на тех наборах переменных, на которых $f = 0$ и $g = 0$. Таким образом, по определению

$$(f \downarrow g)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \cdot g(x_1, \ldots, x_n) + f(x_1, \ldots, x_n) + g(x_1, \ldots, x_n) \pmod 2.$$

Стрелка Пирса называется также *антидизъюнкцией* функций f и g.

1.12 Формулы

Пусть $f(x_1, \ldots, x_n)$ — произвольная булева функция от n переменных. Формулой, представляющей функцию f, называется любое выражение, составленное из символов переменных $x_1, \ldots, x_n$, символов констант 0 и 1, символов функций $\cdot$, $+$, $\to$, $\sim$, $\mid$, $\downarrow$ и скобок, которое при подстановке вместо переменных $x_1, \ldots, x_n$ произвольных булевых функций $g_1, \ldots, g_n$ соответственно превращается в выражение, задающее функцию $f(g_1, \ldots, g_n)$.

Например, выражение $x_1 \cdot x_2 + x_3$ является формулой, представляющей функцию $f(x_1, x_2, x_3) = x_1 x_2 + x_3 \pmod 2$. Подставляя вместо x_1, x_2, x_3 произвольные функции g_1, g_2, g_3, получаем выражение $g_1 g_2 + g_3$, задающее функцию $f(g_1, g_2, g_3)$.

Формулы A и B называются *равносильными*, если они представляют одну и ту же функцию. Равносильность формул A и B обозначается $A = B$.

Например, $x_1 \cdot x_2 + x_3 = x_3 + x_2 \cdot x_1$.

Формула A называется *двойственной* к формуле B, если A получается из B заменой каждого вхождения символа $\cdot$ на $+$, каждого вхождения $+$ на $\cdot$, каждого вхождения 0 на 1, каждого вхождения 1 на 0. Формулу, двойственную к формуле B, обозначают B^*.

Например, $(x_1 \cdot x_2 + x_3)^* = (x_1 + x_2) \cdot x_3$.

Функция $f^*(x_1, \ldots, x_n) = f(x_1 + 1, \ldots, x_n + 1) + 1$ называется *двойственной* к функции $f(x_1, \ldots, x_n)$.

Например, $(x_1 x_2 + x_3)^* = (x_1 + x_2)x_3$.

Функция f называется *самодвойственной*, если $f = f^*$.

Например, функция $x_1 + x_2 + x_3$ самодвойственна.

Функция f называется *монотонной*, если для любых двух наборов переменных $(\alpha_1, \ldots, \alpha_n)$ и $(\beta_1, \ldots, \beta_n)$, таких, что $\alpha_1 \le \beta_1, \ldots, \alpha_n \le \beta_n$, выполняется неравенство $f(\alpha_1, \ldots, \alpha_n) \le f(\beta_1, \ldots, \beta_n)$.

Например, функция $x_1 x_2 + x_3$ монотонна.

Функция f называется *линейной*, если $f = a_0 + a_1 x_1 + \ldots + a_n x_n \pmod 2$, где $a_0, a_1, \ldots, a_n$ — константы, равные 0 или 1.

Например, функция $x_1 + x_2 + x_3$ линейна.

Функция f называется *функцией сохранения нуля*, если $f(0, \ldots, 0) = 0$.

Например, функция $x_1 x_2 + x_3$ сохраняет нуль.

Функция f называется *функцией сохранения единицы*, если $f(1, \ldots, 1) = 1$.

Например, функция $x_1 x_2 + x_3$ не сохраняет единицу.

Функция f называется *существенной* по переменной x_i, если существуют такие наборы переменных $(\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n)$ и $(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n)$, что

$$f(\alpha_1, \ldots, \alpha_{i-1}, 0, \alpha_{i+1}, \ldots, \alpha_n) \ne f(\beta_1, \ldots, \beta_{i-1}, 1, \beta_{i+1}, \ldots, \beta_n).$$

Переменная x_i называется *существенной переменной* функции f, если f существенна по x_i.

Например, функция $x_1 x_2 + x_3$ существенна по x_1, x_2, x_3.

Функция f называется *несущественной* по переменной x_i, если f не существенна по x_i.

Переменная x_i называется *несущественной переменной* функции f, если f несущественна по x_i.

Например, функция $x_1 x_2 + x_3$ несущественна по x_4.

Функции f и g называются *равными*, если они имеют одинаковые наборы существенных переменных и на одинаковых наборах значений существенных переменных принимают одинаковые значения.

Например, функции $x_1 x_2 + x_3$ и $x_2 x_1 + x_3$ равны.

Функции f и g называются *равносильными*, если f получается из g путем добавления или удаления несущественных переменных.

Например, функции $x_1 x_2 + x_3$ и $x_1 x_2 + x_3 + x_4$ равносильны.

Функции f и g называются *тождественными*, если они равны и имеют одинаковые наборы существенных переменных.

Например, функции $x_1 x_2 + x_3$ и $x_2 x_1 + x_3$ тождественны.

Функции f и g называются *различными*, если они не тождественны.

Например, функции $x_1 x_2 + x_3$ и $x_1 x_2 + x_4$ различны.

Функции f и g называются *существенно различными*, если они не равны.

Например, функции $x_1 x_2 + x_3$ и $x_1 x_2 + x_4$ существенно различны.

Функции f и g называются *равносильно различными*, если они не равносильны.

Например, функции $x_1 x_2 + x_3$ и $x_1 x_2 + x_4$ равносильно различны.

Функции f и g называются *тождественно различными*, если они не тождественны.

Например, функции $x_1 x_2 + x_3$ и $x_1 x_2 + x_4$ тождественно различны.

Функции f и g называются *существенно равными*, если они равны.

Например, функции $x_1 x_2 + x_3$ и $x_2 x_1 + x_3$ существенно равны.

Функции f и g называются *равносильно равными*, если они равносильны.

Например, функции $x_1 x_2 + x_3$ и $x_1 x_2 + x_3 + x_4$ равносильно равны.

Функции f и g называются *тождественно равными*, если они тождественны.

Например, функции $x_1 x_2 + x_3$ и $x_2 x_1 + x_3$ тождественно равны.

Функция f называется *функцией алгебры логики*, если f принимает значения 0 или 1 на наборах значений переменных, состоящих из 0 и 1.

Например, функция $x_1 x_2 + x_3$ является функцией алгебры логики.

Функция f называется *функцией многозначной логики*, если f принимает значения 0, 1, 2, ..., $k-1$ на наборах значений переменных, состоящих из 0, 1, 2, ..., $k-1$.

Например, функция $x_1 x_2 + x_3$ является функцией многозначной логики.

Функция f называется *функцией счетнозначной логики*, если f принимает значения 0, 1, 2, ... на наборах значений переменных, состоящих из 0, 1, 2, ...

Например, функция $x_1 x_2 + x_3$ является функцией счетнозначной логики.

Функция f называется *функцией континуальнозначной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел.

Например, функция $x_1 x_2 + x_3$ является функцией континуальнозначной логики.

Функция f называется *функцией вероятностной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся вероятностями.

Например, функция $x_1 x_2 + x_3$ является функцией вероятностной логики.

Функция f называется *функцией нечеткой логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся степенями принадлежности.

Например, функция $x_1 x_2 + x_3$ является функцией нечеткой логики.

Функция f называется *функцией интервальной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся интервалами.

Например, функция $x_1 x_2 + x_3$ является функцией интервальной логики.

Функция f называется *функцией комплексной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся комплексными числами.

Например, функция $x_1 x_2 + x_3$ является функцией комплексной логики.

Функция f называется *функцией кватернионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся кватернионами.

Например, функция $x_1 x_2 + x_3$ является функцией кватернионной логики.

Функция f называется *функцией октавной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся октавами.

Например, функция $x_1 x_2 + x_3$ является функцией октавной логики.

Функция f называется *функцией седионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся седионами.

Например, функция $x_1 x_2 + x_3$ является функцией седионной логики.

Функция f называется *функцией трионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся трионами.

Например, функция $x_1 x_2 + x_3$ является функцией трионной логики.

Функция f называется *функцией дуонной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся дуонами.

Например, функция $x_1 x_2 + x_3$ является функцией дуонной логики.

Функция f называется *функцией гексадекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся гексадекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией гексадекарионной логики.

Функция f называется *функцией тридекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся тридекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией тридекарионной логики.

Функция f называется *функцией эннеадекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся эннеадекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией эннеадекарионной логики.

Функция f называется *функцией пентадекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся пентадекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией пентадекарионной логики.

Функция f называется *функцией гексадекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся гексадекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией гексадекарионной логики.

Функция f называется *функцией тридекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся тридекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией тридекарионной логики.

Функция f называется *функцией эннеадекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся эннеадекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией эннеадекарионной логики.

Функция f называется *функцией пентадекарионной логики*, если f принимает значения на наборах значений переменных, состоящих из чисел, и f принимает значения, являющиеся пентадекарионами.

Например, функция $x_1 x_2 + x_3$ является функцией пентадекарионной логики.

The idea is clearly similar to that underlying the critical region construction of prediction confidence regions (Section 2.2). The extension to general families of distributions is obvious.

From the form of the binomial result (3.1) it is evident that Fisher's definition will not produce the type of factorization (2.6) for Bayes posterior distributions. (Of incidental interest in this regard is the fact that Fisher [9 , p.393] implicitly claimed that such a factorization was impossible.) However, the definition of FLIK leads to unsatisfactory results at a more primitive level, as we shall now show.

One fairly reasonable requirement of predictive likelihood is that infinite amounts of sample data should lead to the actual probability density of the future variable Z. Of course this is not a logical requirement. In the binomial example, it is quite easy to see that as $n \to \infty$, with $y/n \to \theta$,

$$\mathrm{FLIK}(z|y) \to \frac{m^m}{z^z(m-z)^{m-z}}\ \theta^z(1-\theta)^{m-z} ,$$

which is the probability density at z divided by its supremum over the parameter space. This is easy to prove for general regular distributions of Y and Z, and so leads to a straightforward modification of Fisher's definition

$$\mathrm{FLIK}^*(z|y) = \frac{\sup_\theta \mathrm{lik}_{Y,Z}(\theta;\, y,z)}{\sup_\theta \mathrm{lik}_Y(\theta;y)} . \qquad (3.2)$$

This does not resolve matters completely, however, since a second, even more appealing, requirement is that if in the binomial case $m \to \infty$ with $z/m \to \psi$, then the predictive likelihood should converge to the ordinary likelihood of ψ given $Y = y$. It is a simple exercise to show that neither (3.1) nor the modification (3.2) possess this property.

Further discussion of Fisher's predictive likelihood may be found in an essay on the "Fundamental Problem" by Edwards [4]. The modification (3.2) has been studied by Faulkenberry and Lejeune (unpublished report, Oregon State University).

3.3 A New Predictive Likelihood

The basic idea behind Fisher's proposal was that predictive likelihood should reflect the degree to which the values y and z of the observed and unobserved variables support the true hypothesis of common distribution. This idea can be applied in a different way in certain simplifying circumstances, namely when the distributions involved belong to exponential families.

Suppose that $Y_1,\ldots,Y_N$, $N = n + n_F$, are i.i.d. random variables each with p.d.f. $f_Y(y; \theta)$, $\theta \in \Omega$, such that $Y = (Y_1,\ldots,Y_n)$ is to be observed and $Z = (Y_{n_P+1},\ldots,Y_N)$ is to be predicted. Further suppose that $s(\cdot)$ is the minimal sufficient reduction function, and denote $s(Y) = s_P$, $s(Z) = s_F$, $s(Y,Z) = s_{PF}$. Then we propose to measure the likelihood of z by the plausibility of the observed value y given s_{PF}. This notion is exactly that used in defining parametric likelihood, where θ is substituted for s_{PF}. Specifically, the likelihood is to be proportional to the conditional density of $Y = y$ given $S_{PF} = s(y,z)$. A "surprising" or "unlikely" value of z will be one which would a posteriori make the sample value y appear improbable. Since the conditional density of Y given S_{PF} factorizes into $f_{Y|S_P}(y|s(y))f_{S_P|S_{PF}}(s(y)|s(y,z))$, only the second factor is of interest, so that we have

Definition 1 For i.i.d. random variables $Y_1,\ldots,Y_N$, $N = n_P + n_F$, such that $S_P = s(Y_1,\ldots,Y_{n_P})$ and $S_{PF} = s(Y_1,\ldots,Y_N)$ are minimal sufficient reductions, the predictive likelihood of $S_F = s(Z) = s(Y_{n_P+1},\ldots,Y_N)$ is

$$\mathrm{PLIK}(s_F|s_P) = f_{S_P|S_{PF}}(s(y)|s(y,z)) \ . \tag{3.3}$$

$$\mathrm{MLLR}(s_q^{\prime}|s_q) = \frac{1}{q!}\sum_{\pi(\cdot)} p(\pi(\cdot)|s_q)\,\mathrm{MLLR}(s_{\pi(1)}^{\prime}|s_1)\cdots\mathrm{MLLR}(s_{\pi(q)}^{\prime}|s_q) \tag{6.6}$$

where $\pi(\cdot)$ ranges over all permutations of $\{1,\ldots,q\}$ and $p(\pi(\cdot)|s_q)$ is the probability that $(X_1,\ldots,X_q)$ belongs to the cell $(X_{\pi(1)},\ldots,X_{\pi(q)})$ of the partition π_q of $\mathcal{X}$ associated with s_q. Thus, the MLLR of $s_q^{\prime}$ given s_q is the average of the MLLR's of all the pairs $(s_{\pi(1)}^{\prime}|s_1),\ldots,(s_{\pi(q)}^{\prime}|s_q)$ corresponding to the same partition π_q of $\mathcal{X}$ but to different permutations $\pi(\cdot)$ of $\{1,\ldots,q\}$. Since there are $q!$ such permutations, the factor $1/q!$ in (6.6) is necessary to avoid counting the same partition $q!$ times.

Let us now consider the case where s_q is the q-th moment of a probability measure P on $\mathcal{X}$ and $s_q^{\prime}$ is the q-th moment of the image $P^{\prime}$ of P under a measurable transformation T of $\mathcal{X}$ into itself. Then $s_q = \int x^q P(dx)$ and $s_q^{\prime} = \int x^q P^{\prime}(dx) = \int T(x)^q P(dx)$. Hence, if $T(x) = \sum_{i=0}^{r} a_i x^i$ is a polynomial of degree r, then $s_q^{\prime}$ is a linear combination of $s_0, s_1, \ldots, s_{qr}$. Conversely, if T is a measurable transformation of $\mathcal{X}$ into itself such that $s_q^{\prime}$ is a linear combination of $s_0, s_1, \ldots, s_{qr}$ for $q = 1, 2, \ldots$, then it is easy to show that T must be a polynomial of degree $\leq r$. Thus, the MLLR of $s_q^{\prime}$ given s_q is finite if and only if T is a polynomial of degree $\leq r$.

Let us now consider the case where s_q is the q-th moment of a probability measure P on $\mathcal{X}$ and $s_q^{\prime}$ is the q-th moment of the image $P^{\prime}$ of P under a measurable transformation T of $\mathcal{X}$ into itself. Then $s_q = \int x^q P(dx)$ and $s_q^{\prime} = \int x^q P^{\prime}(dx) = \int T(x)^q P(dx)$. Hence, if $T(x) = \sum_{i=0}^{r} a_i x^i$ is a polynomial of degree r, then $s_q^{\prime}$ is a linear combination of $s_0, s_1, \ldots, s_{qr}$. Conversely, if T is a measurable transformation of $\mathcal{X}$ into itself such that $s_q^{\prime}$ is a linear combination of $s_0, s_1, \ldots, s_{qr}$ for $q = 1, 2, \ldots$, then it is easy to show that T must be a polynomial of degree $\leq r$. Thus, the MLLR of $s_q^{\prime}$ given s_q is finite if and only if T is a polynomial of degree $\leq r$.

6.3 A new theorem concerning admissibility

Let us now consider the case where s_q is the q-th moment of a probability measure P on $\mathcal{X}$ and $s_q^{\prime}$ is the q-th moment of the image $P^{\prime}$ of P under a measurable transformation T of $\mathcal{X}$ into itself. Then $s_q = \int x^q P(dx)$ and $s_q^{\prime} = \int x^q P^{\prime}(dx) = \int T(x)^q P(dx)$. Hence, if $T(x) = \sum_{i=0}^{r} a_i x^i$ is a polynomial of degree r, then $s_q^{\prime}$ is a linear combination of $s_0, s_1, \ldots, s_{qr}$. Conversely, if T is a measurable transformation of $\mathcal{X}$ into itself such that $s_q^{\prime}$ is a linear combination of $s_0, s_1, \ldots, s_{qr}$ for $q = 1, 2, \ldots$, then it is easy to show that T must be a polynomial of degree $\leq r$. Thus, the MLLR of $s_q^{\prime}$ given s_q is finite if and only if T is a polynomial of degree $\leq r$.

Of course this likelihood is independent of θ since S_{PF} is sufficient. Also, given S_F, the probability distribution of Z is known, independent of θ, so that only the value of s_F is of concern to us.

Certain interesting features of the definition are clearly illustrated in the case of the "Fundamental Problem".

Example 3.1 Let $Y_1,\ldots,Y_N$ be i.i.d. Bernoulli variables with

$$PR(Y_j = 1) = \theta = 1 - PR(Y_j = 0) \quad .$$

Then if n_P and n_F are fixed, $S_P = \sum_{j=1}^{n} Y_j$, $S_F = \sum_{j=1}^{n_F} Y_{n_P+k}$ and $S_F = S_P + S_F$. Simple calculation of the conditional probability in Definition 1 gives

$$PLIK(s_F|s_P) = \binom{n_P}{s_P}\binom{n_F}{s_F} \div \binom{n_P + n_F}{s_P + s_F}, \tag{3.4}$$

which is the hypergeometric likelihood for sampling without replacement n_P times from an urn containing $s_P + s_F$ "ones" and $N - (s_P + s_F)$ "zeros." The special case $n_F = 1$ gives likelihood odds

$$\frac{PLIK(1|s_P)}{PLIK(0|s_P)} = \frac{s_P + 1}{n_P - s_P + 1} \quad ,$$

formally the same as Laplace's Law of Succession.

The consistency properties described in Section 3.2 are clearly satisfied by (3.4), i.e.

(i) if $n_P \to \infty$ such that $n_P^{-1}s_P \to \theta$, n_F fixed,

$$\lim_{n_P\to\infty} PLIK(s_F|s_P) = \binom{n_F}{s_F} \theta^{s_F} (1 - \theta)^{n_F - s_F} = PR(S_F = s_F|\theta); \tag{3.5}$$

convergence here is w.p.1, at the rate $n_P^{-\frac{1}{2}}$.

(ii) if $n_F \to \infty$ such that $n_F^{-1}s_F \to \psi$, n_P fixed,

$$\lim_{n_F\to\infty} PLIK(s_F|s_P) = \binom{n_P}{s_P} \psi^{s_P} (1 - \psi)^{n_P - s_P} = lik_{s_P}(\psi|s_P, n_P). \tag{3.6}$$

□□□

□□□

(6.3) $\quad \mathrm{MLE}(\theta_q|\theta_q) = \max_{\theta_1,\dots,\theta_{q-1}} \mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q) = \binom{n}{q} \theta_q^q (1-\theta_q)^{n-q}$,

.$\theta_q = \frac{q}{n}$ ביצד $\theta_1,\dots,\theta_{q-1}$ לגבי $\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q)$ את מקסם

<u>הוכחה</u>: נשים לב ש- $\theta_1,\dots,\theta_q$ הם פרמטרים של ההתפלגות $B(n,\theta_q)$, ולכן

$$\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q) = \prod_{i=1}^{n} P(X_i=\theta_q|\theta_1,\dots,\theta_q) = \prod_{i=1}^{n} P(X_i=\theta_q|\theta_q) = \theta_q^{\sum_{i=1}^{n} X_i} (1-\theta_q)^{n-\sum_{i=1}^{n} X_i} = \theta_q^{q} (1-\theta_q)^{n-q} .$$

מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא הפתרון היחיד למשוואה $\frac{\partial}{\partial \theta_q} \mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q) = 0$ בקטע $(0,1)$, ולכן הוא נקודת מקסימום מקומי יחידה של $\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q)$ בקטע זה. כמו כן, $\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q) = 0$ עבור $\theta_q = 0$ ועבור $\theta_q = 1$, ולכן $\theta_q = \frac{q}{n}$ הוא נקודת מקסימום גלובלית יחידה של $\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q)$ בקטע $[0,1]$. מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא הפתרון היחיד למקסם את $\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q)$ ביחס ל- $\theta_1,\dots,\theta_q$.

נשים לב כי $\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q)$ הוא פונקציה של θ_q בלבד, ולכן $\theta_q = \frac{q}{n}$ הוא גם הפתרון היחיד למקסם את $\mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q)$ ביחס ל- θ_q בלבד. כלומר,

$$\max_{\theta_q} \mathrm{MLE}(\theta_q|\theta_q) = \max_{\theta_1,\dots,\theta_{q-1}} \mathrm{MLE}(\theta_1,\dots,\theta_q|\theta_q) = \binom{n}{q} \theta_q^q (1-\theta_q)^{n-q} = \mathrm{MLE}(\theta_q|\theta_q) \quad : (6.2)$$

.$\theta_q = \frac{q}{n}$ כאשר $\max_{\theta_q} \mathrm{MLE}(\theta_q|\theta_q)$ מתקבל

(ב) נוכיח כי $\theta_q = \frac{q}{n}$ הוא גם פתרון של בעיית המקסימום $\max_{\theta_q} \mathrm{MLE}(\theta_q|\theta_q)$ בקטע $[0,1]$.

נשים לב כי $\mathrm{MLE}(\theta_q|\theta_q)$ הוא פונקציה רציפה בקטע $[0,1]$, ולכן היא מקבלת מקסימום בקטע זה. נבדוק את נקודות הקיצון של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $(0,1)$:

$$\frac{\partial}{\partial \theta_q} \mathrm{MLE}(\theta_q|\theta_q) = \binom{n}{q} \left[q\theta_q^{q-1}(1-\theta_q)^{n-q} - (n-q)\theta_q^{q}(1-\theta_q)^{n-q-1} \right] = \binom{n}{q} \theta_q^{q-1}(1-\theta_q)^{n-q-1} \left[q(1-\theta_q) - (n-q)\theta_q \right] = \binom{n}{q} \theta_q^{q-1}(1-\theta_q)^{n-q-1} (q - n\theta_q) .$$

מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא נקודת הקיצון היחידה של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $(0,1)$. מכיוון ש- $\mathrm{MLE}(\theta_q|\theta_q) = 0$ עבור $\theta_q = 0$ ועבור $\theta_q = 1$, נובע כי $\theta_q = \frac{q}{n}$ הוא נקודת מקסימום גלובלית יחידה של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $[0,1]$. מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא הפתרון היחיד למקסם את $\mathrm{MLE}(\theta_q|\theta_q)$.

(ג) נוכיח כי $\theta_q = \frac{q}{n}$ הוא גם פתרון של בעיית המקסימום $\max_{\theta_q} \mathrm{MLE}(\theta_q|\theta_q)$ בקטע $[0,1]$.

נשים לב כי $\mathrm{MLE}(\theta_q|\theta_q)$ הוא פונקציה רציפה בקטע $[0,1]$, ולכן היא מקבלת מקסימום בקטע זה. נבדוק את נקודות הקיצון של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $(0,1)$:

$$\frac{\partial}{\partial \theta_q} \mathrm{MLE}(\theta_q|\theta_q) = \binom{n}{q} \left[q\theta_q^{q-1}(1-\theta_q)^{n-q} - (n-q)\theta_q^{q}(1-\theta_q)^{n-q-1} \right] = \binom{n}{q} \theta_q^{q-1}(1-\theta_q)^{n-q-1} \left[q(1-\theta_q) - (n-q)\theta_q \right] = \binom{n}{q} \theta_q^{q-1}(1-\theta_q)^{n-q-1} (q - n\theta_q) .$$

מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא נקודת הקיצון היחידה של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $(0,1)$. מכיוון ש- $\mathrm{MLE}(\theta_q|\theta_q) = 0$ עבור $\theta_q = 0$ ועבור $\theta_q = 1$, נובע כי $\theta_q = \frac{q}{n}$ הוא נקודת מקסימום גלובלית יחידה של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $[0,1]$. מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא הפתרון היחיד למקסם את $\mathrm{MLE}(\theta_q|\theta_q)$.

(ד) נוכיח כי $\theta_q = \frac{q}{n}$ הוא גם פתרון של בעיית המקסימום $\max_{\theta_q} \mathrm{MLE}(\theta_q|\theta_q)$ בקטע $[0,1]$.

נשים לב כי $\mathrm{MLE}(\theta_q|\theta_q)$ הוא פונקציה רציפה בקטע $[0,1]$, ולכן היא מקבלת מקסימום בקטע זה. נבדוק את נקודות הקיצון של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $(0,1)$:

$$\frac{\partial}{\partial \theta_q} \mathrm{MLE}(\theta_q|\theta_q) = \binom{n}{q} \left[q\theta_q^{q-1}(1-\theta_q)^{n-q} - (n-q)\theta_q^{q}(1-\theta_q)^{n-q-1} \right] = \binom{n}{q} \theta_q^{q-1}(1-\theta_q)^{n-q-1} \left[q(1-\theta_q) - (n-q)\theta_q \right] = \binom{n}{q} \theta_q^{q-1}(1-\theta_q)^{n-q-1} (q - n\theta_q) .$$

מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא נקודת הקיצון היחידה של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $(0,1)$. מכיוון ש- $\mathrm{MLE}(\theta_q|\theta_q) = 0$ עבור $\theta_q = 0$ ועבור $\theta_q = 1$, נובע כי $\theta_q = \frac{q}{n}$ הוא נקודת מקסימום גלובלית יחידה של $\mathrm{MLE}(\theta_q|\theta_q)$ בקטע $[0,1]$. מכאן נובע כי $\theta_q = \frac{q}{n}$ הוא הפתרון היחיד למקסם את $\mathrm{MLE}(\theta_q|\theta_q)$.

It is clear from Definition 1 that its main domain is that of exponential family distributions, i.e. in the i.i.d. case distributions with densities of the form

$$f_{Y_j}(y|\theta) = \exp\{-\theta' b(y) + c(\theta) + d(y)\} \tag{3.7}$$

where θ, possibly a vector, is unrestricted. Some further comment will be given on this later. It is not hard to see that consistency properties of the type (3.5) and (3.6) will hold generally for such exponential families. Indeed, the simple relationship between the sufficient statistic $s = \Sigma b(y_j)$ and the maximum likelihood estimate $\hat{\theta}$ for (3.7) enables one to show that, in obvious notation,

$$\text{PLIK}(s_F|s_P) = \exp\{-\hat{\theta}_P' S_F + n_F c(\hat{\theta}_P) + d^*(s_F)\} + O(n_P^{-1}). \tag{3.8}$$

Some details are given in the Appendix.

Before continuing with the general development of Definition 1, it is of interest to consider the inferential use of PLIK. According to the discussion of Section 2.2, this new likelihood would be used to remove non-uniqueness of confidence regions by choosing a region $P_\alpha(y)$ such that

$$\sup_{z \notin P_\alpha(y)} \text{PLIK}(s_F(z)|s_p(y)) \leq \inf_{z \in P_\alpha(y)} \text{PLIK}(s_F(z)|s_p(y)). \tag{3.9}$$

This would seem to be a non-controversial proposal. Now recall the other difficulty with critical region construction of confidence sets, namely the possibility that the hypothesis testing problem does not possess Neyman structure. In certain cases, typified by the following example, a controversial application of PLIK can overcome this difficulty, essentially by treating PLIK as a density.

It is clear from Definition 1 that its main domain is that of exponential family distributions, i.e. in the i.i.d. case distributions with densities of the form

$$f_\theta^n(x/\theta) = \exp\{-n[b(x) + c(\theta) + d(x)]\} \qquad (3.1)$$

where θ, possibly a vector, is unrestricted. Some further comment will be given on this later. It is not hard to see that coincidence properties of the type (3.3) and (3.6) will hold generally for such exponential families. Indeed, the existence relationship between the multichart structure $s = \mathcal{D}(\tau_\beta^\alpha)$ and the maximum likelihood estimate $\hat\theta$ for (3.7) enables one to show that, in obvious notation,

$$\mathrm{PMIX}(s_\beta^\alpha / s_\beta) = \exp\{-n\hat\theta_\beta^\alpha\} = \exp\{-n[c(\hat\theta_\beta) + d(x_\beta)]\} + O(n^{-1}). \qquad (3.8)$$

Some details are given in the Appendix.

Before continuing with the mathematical development of Definition 1, it is of interest to consider the theoretical use of PMIX. According to the discussion of Section 2.2, this new likelihood would be used to remove non-uniqueness of coincidence regions by choosing a region $R_\beta^\alpha(x)$ such that

$$\sup_{s_\beta^\alpha(x)} \mathrm{PMIX}(s_\beta^\alpha(x) | s_\beta(x)) \le \inf_{s_\beta^\alpha(x)} \mathrm{PMIX}(s_\beta^\alpha(x) | s_\beta(x)). \qquad (3.9)$$

This would seem to be a non-controversial proposal. How would the option of uniformity with critical region construction of confidence sets, namely the possibility that the hypothesis testing version does not possess Neyman structure. In certain cases, typified by the following example, a controversial application of PMIX can overcome this difficulty, essentially by treating PMIX as a density.

Example 3.2 Suppose that $Y_1,\ldots,Y_n$ are counts in non-overlapping unit time intervals of a Poisson process with constant rate θ, and let Y_t^* be the corresponding count in a future interval of length t. Application of (3.3) easily shows that with $S = \Sigma Y_j$

$$\text{PLIK}(y^*|s) = \frac{\Gamma(s + y^* + 1)}{\Gamma(s + 1)} n^s (n + t)^{s+y^*}. \qquad (3.10)$$

But now suppose that we wish to make inference about the time interval T from the present to the occurrence of the next event; note that neither of the sampling-theory approaches of Section 2.2 could deal with this problem exactly. We observe that, in the earlier notation, $T \geq t$ if and only if $Y_t^* = 0$. Therefore, the predictive likelihood of the event "$T \geq t$" is precisely (3.10) evaluated at $y^* = 0$. It is then tempting to take

$$\text{PLIK}(t|s) = -\frac{\partial}{\partial t} \text{PLIK}(0|s) = sn^s (n + t)^{-s-1} , \qquad (3.11)$$

although there is no logical case for this manipulation. Notice that the result satisfies the natural consistency property

$$\lim_{n\to\infty} \text{PLIK}(t|S) = \theta\exp(-\theta t) \quad \text{w.p.1} .$$

Not surprisingly, (3.11) is slightly different to the analogous (logical) result for inter-event time observations; in particular, for total time n between the start of observation and occurrence of the s^{th} event, with s fixed,

$$\text{PLIK}(t|n) = sn^{s-1}(n + t)^{-s} . \qquad \square\square\square$$

Similar manipulations are possible in several other problems where the observable Y and future variable Z are determined by different sampling

Lemma 2.2. Suppose that $X_1, X_2, \ldots$ are common m non-negative independent random variables with a distribution function F such that $F(0) < 1$, and let $\hat{F}$ be the corresponding Laplace-Stieltjes transform of F. Then the renewal function H has a derivative h which is completely monotone, and h is the unique solution of the renewal equation

$$h(x) = f(x) + \int_0^x h(x-y)\,dF(y) \qquad (x \geq 0)$$

with $f(x) = e^{-\lambda x}$, where λ is the unique positive root of the equation $\hat{F}(\lambda) = 1$.

Proof. Let $\{N(t), t \geq 0\}$ be the renewal counting process corresponding to the X_i's, and let $\{Z(t), t \geq 0\}$ be the associated forward recurrence time process. Then for $x > 0$ and $t > 0$,

$$P\{Z(t) > x \mid Z(0) = 0\} = P\{N(t+x) - N(t) = 0 \mid N(0) = 0\} = 1 - F(t+x) + \int_0^t P\{Z(t-y) > x \mid Z(0) = 0\}\,dF(y).$$

Hence, by Theorem 2.1, $P\{Z(t) > x \mid Z(0) = 0\}$ converges to $1/\mu$ times the integral from x to ∞ of $1 - F(y)$, where $\mu = \int_0^\infty y\,dF(y)$ is the mean of F. Since $P\{Z(t) > x \mid Z(0) = 0\} = 1 - H(t) + \int_0^t [1 - H(t-y)]\,dF(y)$, it follows that $H(t)$ converges to $1/\mu$ times the integral from 0 to t of $1 - F(y)$. Since $F(0) < 1$, $\mu > 0$, and hence $H(t)$ converges to a finite limit as $t \to \infty$. Therefore, by Theorem 2.2, $h(t)$ converges to 0 as $t \to \infty$.

Now, by the renewal theorem, $h(t)$ satisfies the renewal equation

$$h(t) = f(t) + \int_0^t h(t-y)\,dF(y) \qquad (t \geq 0), \qquad (2.4)$$

where $f(t) = F(t) + \int_0^t f(t-y)\,dF(y)$. Since $f(t)$ is completely monotone, it follows from Theorem 2.3 that $h(t)$ is completely monotone if and only if $f(t) = e^{-\lambda t}$ for some $\lambda \geq 0$. Substituting $e^{-\lambda t}$ for $f(t)$ in (2.4) and dividing by $e^{-\lambda t}$ yields

$$1 = \hat{F}(\lambda) + \int_0^t e^{\lambda y}\,dF(y) - \int_t^\infty e^{\lambda(t-y)}\,dF(y).$$

Since $F(0) < 1$, $\hat{F}(\lambda)$ is a strictly increasing function of λ with $\hat{F}(0) = 0$ and $\hat{F}(\lambda) \to \infty$ as $\lambda \to \infty$. Hence there is a unique $\lambda > 0$ such that $\hat{F}(\lambda) = 1$. For this λ, $\int_0^t e^{\lambda y}\,dF(y) = e^{\lambda t} - \int_t^\infty e^{\lambda(t-y)}\,dF(y)$, and hence $h(t) = e^{-\lambda t}$ satisfies (2.4). Finally, $h(t)$ is the unique solution of (2.4) since any two solutions must differ by a multiple of $e^{-\lambda t}$, and no such multiple can be a solution unless it is identically zero.

□□□

Another method of finding the renewal function H is to solve the renewal equation (2.4) numerically. Methods for doing this are discussed in Cox and Miller [1965, pp. 150-152]. Another method is to approximate H by the distribution function of a mixture of Erlang distributions. Tocher [1963, pp. 75-77] has shown how this can be done.

rules. The full consequences of such a wide interpretation of predictive likelihood have not been investigated.

To return to general developments, we now consider an extension of Definition 1 that is applicable in the non-i.i.d. exponential family case. It is not hard to see that in order for (3.3) to give the desired result for Z, independent of θ, we require only that

(i) the conditional distribution of Z given s_F and s_P be known independent of θ

(ii) $S_{PF} = s^*(S_P, S_F)$ for some function $s^*(\cdot)$

and (iii) $f_{s_{PF}}(s_{PF};\theta) = k(s_P, s_F) f_{s_P}(s_P;\theta) f_{s_F|s_P}(s_F|s_P;\theta)$. (3.12)

Thus, in general, it is not <u>necessary</u> for s_F to be a sufficient reduction of Z, nor is it <u>sufficient</u> for s_P to be a sufficient reduction of Y; see Examples 3.3 and 3.4 below. We therefore make the revised

<u>Definition 2</u> If s_P and s_F are minimal reductions of Y and Z such that (i) - (iii) in (3.12) are satisfied, with s_{PF} the minimal sufficient reduction of (Y, Z), then the predictive likelihood of s_F is

$$PLIK(s_F|s_P) = f_{s_P|s_{PF}}(s_P(y)|s_{PF}(y,z)) \tag{3.13a}$$

and the predictive likelihood of z is

$$PLIK_Z(z|s_P) = f_{Z|s_F,s_P}(z|s_F,s_P)\ PLIK(s_F|s_P) . \tag{3.13b}$$

(The statistic s_F, which depends on s_P, is called a minimal necessary statistic.) Two technical points to note are, firstly, that the condition (ii) can be satisfied if s_P is a minimal totally sufficient statistic as defined by Lauritzen [14]; and, secondly, Definition 2 extends the domain of non-trivial predictive likelihood beyond exponential

отображения по подпоследовательности дают в точке x_0 соприкасающееся
линейное. Для этого достаточно, чтобы все локальные
представления отображения f в окрестности точки x_0 имели
вид $(x_1, \ldots, x_n) \to (x_1, x_2^2, x_3, \ldots, x_n)$.

Из теоремы 1 следует, что множество $\Sigma^1(f)$ особых точек
отображения f является гладким многообразием размерности
$n-1$. Ограничение f на $\Sigma^1(f)$ также является гладким
отображением. Его особые точки образуют множество $\Sigma^2(f)$.
Аналогично определяются $\Sigma^3(f), \ldots, \Sigma^k(f)$. Если
$x \in \Sigma^k(f) \setminus \Sigma^{k+1}(f)$, то ранг df в точке x равен $n-k$.
Поэтому $\Sigma^k(f)$ — гладкое многообразие размерности $n-k$.
Если $k > n$, то $\Sigma^k(f) = \emptyset$.

Пусть $f: M^n \to N^p$ — гладкое отображение гладких
многообразий. Точка $x \in M^n$ называется *точкой общего положения*
отображения f, если существует такая окрестность U точки x,
что для любого k множество $\Sigma^k(f) \cap U$ является
подмногообразием в U размерности $n-k$ или пусто.

Теорема 2 (Уитни). *Для любого гладкого отображения*
$f: M^n \to N^p$ *существует всюду плотное открытое множество*
$V \subset M^n$, *такое, что ограничение* f *на* V *является*
отображением общего положения.

Доказательство этой теоремы мы опускаем. Заинтересованный
читатель может найти его в книге Уитни [1].

Пусть $f: M^n \to N^p$ — гладкое отображение гладких
многообразий. Точка $x \in M^n$ называется *точкой общего положения*
отображения f, если существует такая окрестность U точки x,
что для любого k множество $\Sigma^k(f) \cap U$ является
подмногообразием в U размерности $n-k$ или пусто. (Отсюда
следует, в частности, что $\Sigma^1(f)$ является гладким
многообразием размерности $n-1$ и $\Sigma^2(f)$ — гладким
многообразием размерности $n-2$.) Если x — точка общего
положения, а $f(x) = y$, то говорят, что f имеет в точке x
особенность типа $(x_1, x_2^2, x_3, \ldots, x_n)$. Если x — точка
общего положения, а $f(x) = y$, то говорят, что f имеет в точке x
особенность типа $(x_1, x_2^3, x_3, \ldots, x_n)$. Если x — точка
общего положения, а $f(x) = y$, то говорят, что f имеет в точке x
особенность типа $(x_1, x_2^2 + x_3^2, x_4, \ldots, x_n)$.

Пусть $f: M^n \to N^p$ — гладкое отображение гладких
многообразий. Точка $x \in M^n$ называется *точкой общего положения*
отображения f, если существует такая окрестность U точки x,
что для любого k множество $\Sigma^k(f) \cap U$ является
подмногообразием в U размерности $n-k$ или пусто. (Отсюда
следует, в частности, что $\Sigma^1(f)$ является гладким
многообразием размерности $n-1$ и $\Sigma^2(f)$ — гладким
многообразием размерности $n-2$.) Если x — точка общего
положения, а $f(x) = y$, то говорят, что f имеет в точке x
особенность типа $(x_1, x_2^2, x_3, \ldots, x_n)$. Если x — точка
общего положения, а $f(x) = y$, то говорят, что f имеет в точке x
особенность типа $(x_1, x_2^3, x_3, \ldots, x_n)$. Если x — точка
общего положения, а $f(x) = y$, то говорят, что f имеет в точке x
особенность типа $(x_1, x_2^2 + x_3^2, x_4, \ldots, x_n)$.

families, albeit not much beyond. The following examples illustrate the above discussion.

Example 3.3 Let $Y_1,\ldots,Y_n$ and $Z = Y_{n+1}$ be i.i.d. with uniform density on $[0, \theta]$. Here the minimal sufficient statistic based on $Y_1,\ldots,Y_m$ is the maximum $Y_{(m,m)}$, so that $S_{PF} = \max(Z, Y_{(n,n)})$. Clearly, $S_P = Y_{(n,n)}$, but not all of Z is required to define S_{PF}. In fact to satisfy the condition (ii) in (3.12) we may take

$$S_F = \begin{cases} 0 & Z \leq Y_{(n,n)} \\ Z & Z > Y_{(n,n)} \end{cases} .$$

A simple probability calculation shows that

$$PLIK(s_F|s_P) = \begin{cases} n(n+1)^{-1} & (s_F = 0) \\ n(n+1)^{-1}s_F^{-1}s_P^{n-1} & (s_F > 0) \end{cases} ,$$

and since

$$f_{Z|S_F,S_P}(z|s_P,s_F) = \begin{cases} s_P^{-1} & (s_F = 0) \\ \delta(z-s_F) & (s_F > 0) \end{cases}$$

we have by (3.13) that

$$PLIK_Z(z|s_P) = \begin{cases} n(n+1)^{-1}s_P^{-1} & (s_F = 0) \\ n(n+1)^{-1}s_F^{-1}s_P^{n-1} & (s_F > 0) \end{cases} .$$

This last result is, in fact, a continuous probability density; notice that the predictive likelihood attached to the event $Z \leq Y_{(n,n)}$, i.e. $n(n+1)^{-1}$, formally agrees with the frequency probability of this event. The results for prediction of $n_F > 1$ variables are obvious generalizations. □□□

Следующий результат дает достаточное условие для того, чтобы $\{X_t\}$ было m-зависимым.

Лемма 3.5. Пусть $\{X_t\}$ — стационарный процесс с конечными моментами второго порядка, $\gamma(\cdot)$ — его автоковариационная функция, $\rho(\cdot)$ — автокорреляционная функция. Если $\gamma(m) = 0$, $\gamma(m+1) = 0$ и $\gamma(m+2) = 0$, то $\{X_t\}$ является m-зависимым.

Доказательство. Пусть $\{X_t\}$ — стационарный процесс с $\gamma(m) = \gamma(m+1) = \gamma(m+2) = 0$. Тогда $\rho(m) = \rho(m+1) = 0$, и из (3.1.3) следует, что $\rho(m-1) = 0$. Аналогично, $\rho(m-2) = 0$. Продолжая этот процесс, получаем, что $\rho(1) = \rho(2) = \dots = \rho(m) = 0$. Следовательно, $\gamma(k) = 0$ для всех $k \geq 1$, и $\{X_t\}$ является m-зависимым. □

Теперь мы можем доказать центральную предельную теорему для m-зависимых последовательностей.

Теорема 3.5. (Центральная предельная теорема для m-зависимых последовательностей) Пусть $\{X_t\}$ — стационарная m-зависимая последовательность с $E(X_t) = \mu$ и $\gamma(k) = \operatorname{Cov}(X_t, X_{t+k})$. Если $v_m = \gamma(0) + 2\sum_{k=1}^{m}\gamma(k) \neq 0$, то

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, v_m),$$

где $\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Доказательство. Положим $S_n = \sum_{t=1}^{n} X_t$. Тогда

$$\operatorname{Var}(S_n) = n\gamma(0) + 2\sum_{k=1}^{m}(n-k)\gamma(k) = n v_m - 2\sum_{k=1}^{m} k\gamma(k).$$

Следовательно,

$$\frac{\operatorname{Var}(S_n)}{n} \to v_m.$$

Применяя центральную предельную теорему для сумм независимых случайных величин к последовательности $\{X_t\}$, получаем требуемый результат. □

Example 3.4 Let $\{Y_j\}$ be a stationary first-order binary Markov sequence, with

$$PR(Y_{j+1} = b | Y_j = a) = \theta_{ab} \qquad (a,b = 0,1) \ . \tag{3.13}$$

Suppose that $y = (y_0, y_1, \ldots, y_{n_P})$ is observed, and that $Z = (Y_{n_P+1}, \ldots, Y_{n_P+n_F})$ is to be predicted. For any connected sequence of variables the minimal sufficient statistic is the initial value plus the matrix of one-step transition frequencies. Thus, if we let $\delta(u) = 1(u = 0)$, $= 0(u \neq 0)$ and define the matrix $\underset{\sim}{M}(r,S)$ by

$$M_{ab}(r,S) = \sum_{j=r}^{s-1} \delta(Y_j - a)\delta(Y_{j+1} - b) \qquad (a,b = 0, 1;\ S \geq r) \ ,$$

we may express the minimal sufficient reductions of Y, Z and (Y,Z) as

$$(Y_0, \underset{\sim}{M}(0,n_P)), \ (Y_{n_P+1}, \underset{\sim}{M}(n_P+1, n_P+n_F)) \text{ and } (Y_0, \underset{\sim}{M}(0, n_P+n_F))$$

respectively. Clearly the first two do not give the last, because

$$\underset{\sim}{M}(0,n_P+n_F) = \underset{\sim}{M}(0,n_P) + \underset{\sim}{M}(n_P+1, n_P+n_F) + \begin{pmatrix} \delta(Y_{n_P})\delta(Y_{n_P+1}) & \delta(Y_{n_P})\delta(1-Y_{n_P+1}) \\ \delta(1-Y_{n_P})\delta(Y_{n_P+1}) & \delta(1-Y_{n_P})\delta(1-Y_{n_P+1}) \end{pmatrix}$$

Thus in order to satisfy condition (ii) in (3.11) we must take

$$s_P = (Y_0, \underset{\sim}{M}(0,n_P), Y_{n_P}), \quad s_F = (Y_{n_P+1}, \underset{\sim}{M}(n_P+1, n_P+n_F)) ;$$

s_P is minimal totally sufficient.

The ensuing calculation of predictive likelihood is algebraically complicated, although the basic results required are given by Whittle [20]. For the special case $n_F = 1$ it is straightforward to show that (3.12) gives

$$PLIK(y_{n_P+1} = b | y_0, \underset{\sim}{m}(0,n), y_n = a) = \frac{m_{ab}(0,n)+1}{\sum_{k=0}^{1} m_{ak}(0,n) + 1} \ .$$

<u>Example 3.4</u> Let $\{Y_i\}$ be a stationary first-order binary Markov process, with

$$\Pr(Y_{i+1} = b \mid Y_i = a) = \theta_{ab} \qquad (a,b = 0,1). \qquad (3.13)$$

Suppose that $y = (y_0, y_1, \ldots, y_{n_p})$ is observed, and that $z = (Y_{n_p+1}, \ldots, Y_{n_p+n_f})$ is to be predicted. For any connected sequence of variables the minimal sufficient statistic is the initial value plus the matrix of one-step transition frequencies. Thus, if we let $\delta(u) = 1 (u = 0), = 0 (u \neq 0)$ and define the matrix $N(r,s)$ by

$$N_{ab}(r,s) = \sum_{j=r}^{s-1} \delta(Y_j - a)\,\delta(Y_{j+1} - b) \qquad (a,b = 0,1;\ s \geq r),$$

we may express the minimal sufficient reductions of Y, Z and (Y,Z) as

$$(Y_0, N(0,n_p)),\ (Y_{n_p+1}, N(n_p+1, n_p+n_f)) \text{ and } (Y_0, N(0, n_p+n_f))$$

respectively. Clearly the first two do not give the last, because

$$N(0,n_p+n_f) = N(0,n_p) + N(n_p+1, n_p+n_f) + \begin{pmatrix} \delta(Y_{n_p})\delta(Y_{n_p+1}) & \delta(Y_{n_p})(1-Y_{n_p+1}) \\ (1-Y_{n_p})\delta(Y_{n_p+1}) & (1-Y_{n_p})(1-Y_{n_p+1}) \end{pmatrix}$$

Thus in order to satisfy condition (ii) in (3.11) we must take

$$s_p = (Y_0, N(0,n_p), Y_{n_p}),\quad s_f = (Y_{n_p+1}, N(n_p+1, n_p+n_f));$$

s_p is minimal totally sufficient.

The ensuing calculation of predictive likelihood is algebraically complicated, although the basic results required are given by Whittle [20]. For the special case $n_f = 1$ it is straightforward to show that (3.12) gives

$$\mathrm{PLik}(Y_{n_p+1} = b \mid y_0, N(0,n), y_n = a) = \frac{N_{ab}(0,n) + 1}{\sum_{k=0}^{1} N_{ak}(0,n) + 1}.$$

This corresponds exactly to (3.4); the data in row 1-a of $\underset{\sim}{m}(0,n)$ is ignored when $y_n = a$, no matter how much information exists to support $\theta_{0b} = \theta_{1b}$. □□□

Although the predictive likelihood of Definition 2 can clearly play a role in sampling-theory confidence set construction, and possibly provide a credibility measure as in Example 3.2, there remains the question of the relation with Bayesian prediction. Recall from (2.5) that if θ is the realized value of a random variable Θ with prior density $p(\theta)$, then

$$f_{S_F|S_P}(s_F|s_P) = \frac{\int f_{S_P|\Theta}(s_P|\theta) f_{S_F|S_P,\Theta}(s_F|s_P,\theta) p(\theta) d\theta}{\int f_{S_P|\Theta}(s_P|\theta) p(\theta) d\theta} .$$

Now if condition (iii) in (3.12) is satisfied, we have immediately that

$$f_{S_F|S_P}(s_F|s_P) = \frac{\{k(s_P,s_F)\}^{-1} \int f_{S_{PF}|\Theta}(s_{PF}|\theta) p(\theta) d\theta}{\int f_{S_P|\Theta}(s_P|\theta) p(\theta) d\theta} \propto \text{PLIK}(s_F|s_P) f^o_{S_{PF}}(s^*(s_P,s_F)), \tag{3.14a}$$

where f^o_S is the marginal prior density of S. In the case of independent exponential family variables, when $s_{PF} = s_P + s_F$ in the natural scale, (3.13a) may be expressed as

$$f_{S_F|S_P}(s_F|s_P) \propto \text{PLIK}(s_F|s_P) f^o_{S_F}(s_F) . \tag{3.14b}$$

These results are precisely the desired analogs of the parametric result (2.6). By the consistency property of PLIK, and a corresponding consistency property of s_F as $n_F \to \infty$, the parametric result is in fact the limiting case of (3.14).

It is, then, possible to obtain posterior predictive distributions by assigning prior probabilities directly to the random variables, rather than to parameters, and using a predictive likelihood; this

Hence corresponds exactly to (3.4) and this ends the proof. □ □ □

Note that $\mathbb{E}_{\mathcal{L}(\cdot)}[\mu_{\mathcal{L}(\cdot)}(B)]$ for $B=A\times\mathbb{R}^{n-1}$ was also the result of Theorem 3.1 when $\mathcal{L}=\mathcal{L}_Y$ was a Lévy process. Indeed, in this case $\mathcal{L}_Y(s)=\mathcal{L}_Y(0)$ a.s. for all $s\in[0,t]$ and $\mathcal{L}_Y(0)$ is a random variable independent of Y. Thus, the conditioning in Theorem 3.1 is superfluous and the theorem reduces to the well-known Lévy-Khintchine formula for infinitely divisible laws. Also, the Lévy measure $\nu(\cdot)$ of Y is given by $\nu(B)=\mathbb{E}[\mu_Y(B)]$, $B\in\mathcal{B}(\mathbb{R}^n)$, $0\notin\bar{B}$, which is finite by (3.4). Moreover, the Lévy-Khintchine exponent of Y is

$$\log\mathbb{E}[e^{i\langle z,Y_t\rangle}]=t\log\mathbb{E}[e^{i\langle z,Y_1\rangle}]=t\int_{\mathbb{R}^n}(e^{i\langle z,x\rangle}-1-i\langle z,x\rangle\mathbb{1}_{\{|x|\le 1\}})\,\nu(dx)+i\langle z,\gamma t\rangle,\qquad z\in\mathbb{R}^n,$$

where $\gamma=\mathbb{E}[Y_1]-\int_{\{|x|>1\}}x\,\nu(dx)$. Now, let $\mathcal{L}(\cdot)$ be a general Lévy process with characteristic triplet (A,ν,γ) and let $\mu_{\mathcal{L}}(\cdot)$ be the corresponding jump measure. Then, by the Lévy-Khintchine formula, we have

$$\mathbb{E}[e^{i\langle z,\mathcal{L}_t\rangle}]=\exp\Big(t\int_{\mathbb{R}^n}(e^{i\langle z,x\rangle}-1-i\langle z,x\rangle\mathbb{1}_{\{|x|\le 1\}})\,\nu(dx)+i\langle z,\gamma t\rangle-\tfrac{1}{2}\langle Az,z\rangle t\Big),\qquad z\in\mathbb{R}^n,$$

and, by the Lévy-Itô decomposition, $\mathcal{L}(\cdot)$ can be represented as

$$\mathcal{L}_t=\int_0^t\int_{\{|x|\le 1\}}x\,\tilde{\mu}_{\mathcal{L}}(ds,dx)+\int_0^t\int_{\{|x|>1\}}x\,\mu_{\mathcal{L}}(ds,dx)+\gamma t+\sigma W_t,\qquad t\ge 0,$$

where $\tilde{\mu}_{\mathcal{L}}(ds,dx)=\mu_{\mathcal{L}}(ds,dx)-\nu(dx)\,ds$ is the compensated jump measure of $\mathcal{L}$, W is a standard Brownian motion in $\mathbb{R}^n$, and $\sigma\in\mathbb{R}^{n\times n}$ is such that $\sigma\sigma^T=A$. Now, let $\mathcal{L}(\cdot)$ be a Lévy process with characteristic triplet (A,ν,γ) and let $\mu_{\mathcal{L}}(\cdot)$ be the corresponding jump measure. Then, by the Lévy-Khintchine formula, we have

$$\mathbb{E}[e^{i\langle z,\mathcal{L}_t\rangle}]=\exp\Big(t\int_{\mathbb{R}^n}(e^{i\langle z,x\rangle}-1-i\langle z,x\rangle\mathbb{1}_{\{|x|\le 1\}})\,\nu(dx)+i\langle z,\gamma t\rangle-\tfrac{1}{2}\langle Az,z\rangle t\Big),\qquad z\in\mathbb{R}^n,$$

and, by the Lévy-Itô decomposition, $\mathcal{L}(\cdot)$ can be represented as

$$\mathcal{L}_t=\int_0^t\int_{\{|x|\le 1\}}x\,\tilde{\mu}_{\mathcal{L}}(ds,dx)+\int_0^t\int_{\{|x|>1\}}x\,\mu_{\mathcal{L}}(ds,dx)+\gamma t+\sigma W_t,\qquad t\ge 0,$$

where $\tilde{\mu}_{\mathcal{L}}(ds,dx)=\mu_{\mathcal{L}}(ds,dx)-\nu(dx)\,ds$ is the compensated jump measure of $\mathcal{L}$, W is a standard Brownian motion in $\mathbb{R}^n$, and $\sigma\in\mathbb{R}^{n\times n}$ is such that $\sigma\sigma^T=A$. Now, let $\mathcal{L}(\cdot)$ be a Lévy process with characteristic triplet (A,ν,γ) and let $\mu_{\mathcal{L}}(\cdot)$ be the corresponding jump measure. Then, by the Lévy-Khintchine formula, we have

$$\mathbb{E}[e^{i\langle z,\mathcal{L}_t\rangle}]=\exp\Big(t\int_{\mathbb{R}^n}(e^{i\langle z,x\rangle}-1-i\langle z,x\rangle\mathbb{1}_{\{|x|\le 1\}})\,\nu(dx)+i\langle z,\gamma t\rangle-\tfrac{1}{2}\langle Az,z\rangle t\Big),\qquad z\in\mathbb{R}^n,$$

and, by the Lévy-Itô decomposition, $\mathcal{L}(\cdot)$ can be represented as

$$\mathcal{L}_t=\int_0^t\int_{\{|x|\le 1\}}x\,\tilde{\mu}_{\mathcal{L}}(ds,dx)+\int_0^t\int_{\{|x|>1\}}x\,\mu_{\mathcal{L}}(ds,dx)+\gamma t+\sigma W_t,\qquad t\ge 0,$$

where $\tilde{\mu}_{\mathcal{L}}(ds,dx)=\mu_{\mathcal{L}}(ds,dx)-\nu(dx)\,ds$ is the compensated jump measure of $\mathcal{L}$, W is a standard Brownian motion in $\mathbb{R}^n$, and $\sigma\in\mathbb{R}^{n\times n}$ is such that $\sigma\sigma^T=A$.

provided Definition 2 is non-trivial for the particular model. In a sense this is not new, since for exponential families the prior density $P(\theta)$ may be obtained from the marginal prior density $f^{o}_{S_{PF}}(s_{PF})$ by inverting a Laplace transform. Thus (3.14) is a re-expression in suggestive terms of a known result.

Curiosity prompts one to ask: for what prior distributions is the (normalized) predictive likelihood a posterior predictive density in familiar problems? By (3.14) it is clearly necessary that S_{PF} have a uniform marginal prior density. The corresponding prior densities on Θ for some commonly-occurring models are given in Table 1. Of course, the results depend on the sampling rule, since the predictive likelihood does; in that sense predictive likelihood has slightly lower status than parametric likelihood. However, if you pose a silly question, you get a silly answer.

[Table 1 here]

3.4 Conditional Predictive Likelihood

The predictive likelihood of Definition 2 gives a non-trivial result only when sufficiency arguments produce reductions of Y and Z. Thus the definition is meaningful for unrestricted exponential families and not much further. This leaves a large number of models. However, in many important problems the minimal sufficient statistic, S, although of dimension n, can be expressed as (T,C) where C is an ancillary statistic and T often has the same dimension as θ. Conventionally inference would then proceed conditionally with the value of C regarded as fixed; see [3 , Section 2.3] and [1], for example. In such cases it is possible to obtain a meaningful predictive likelihood by conditioning.

predictive definition y is non-trivial for the particular model. In a sense this is not new, since for exponential families the prior family p(θ) may be obtained from the marginal prior density $\pi_c^{a^{b\theta}}(a^{b\theta})$ by involving a Laplace transform. Thus (3.12) is a re-expression in suggestive terms of a known result.

Conversely, perhaps one of the most useful prior distributions is the (normalised) predictive likelihood a posterior predictive density in familiar problems; by (3.12) it is clearly necessary that e^{bt} have a uniform marginal prior density. The conjugate prior densities of for some commonly-occurring models are given in Table 1. Of course, the results depend on the defining data, since the predictive likelihood does: in that sense predictive likelihood has slightly fewer options than parametric likelihood. However, if you pose a silly question, you get a silly answer.

[Table 1 here]

3.4 Conditional predictive likelihood

The predictive likelihood of statistic y given a non-trivial result only when sufficiency arguments produce reductions of y and z. Thus the definition is meaningful for unrestricted exponential families and not much further. This leaves a large number of models. However, in many important problems the minimal sufficient statistic, s^b, already of dimension n, can be expressed as (t,a) where a is an ancillary statistic and t often has the same dimension as θ. Consequently, inference would then proceed conditionally upon the value of a observed as with; see Cox and Hinkley (1974, section), for example. In such cases it is possible to obtain a meaningful predictive likelihood by conditioning.

Suppose that conditions (i) - (iii) of (3.12) hold with S_P, S_F and S_{PF} minimal. Also suppose that $S_\alpha = (T_\alpha, C_\alpha)$ with C_α ancillary for $\alpha = P, F, PF$. Clearly, C_{PF} is a function of (C_P, C_F, T_P, T_F) involving an ancillary function of (T_P, T_F), and T_{PF} is a function of (T_P, T_F). The statistic C_P contains no information about θ, and hence none about S_F, so we condition on C_P and make

Definition 3 Under the above conditions,

$$\text{PLIK}(s_F|s_P) = f_{C_F}(c_F) f_{T_P|S_{PF},C_P}(t_P|s^*(s_P,s_F),c_P) \tag{3.15}$$

$$= f_{C_F}(c_F)\text{PLIK}_C(t_F|s_P,c_F) \ .$$

When C_{PF} is null, this reduces to Definition 2.

The following familiar example illustrates this definition.

Example 3.5 Let $Y_1,\ldots,Y_N$ be i.i.d. with continuous p.d.f. $g(y-\theta)$, supported on the whole real line, with $Y = (Y_1,\ldots,Y_{n_P})$ and $Z = (Y_{n_P+1},\ldots,Y_{n_P+n_F})$, $N = n_P + n_F$. Denote the ordered values of Y, Z, and (Y,Z) by $\{Y_{(\alpha j)}, j=1,\ldots,n_\alpha\}$ with $\alpha = P$, F, and PF respectively. Then in general $S_\alpha = \{Y_{(\alpha j)}\}$ and we may take

$$T_\alpha = n_\alpha^{-1} \Sigma\, Y_{(\alpha j)}, \quad C_\alpha = \{Y_{(\alpha,j+1)} - Y_{(\alpha,j)},\ j=1,\ldots,n_\alpha-1\}$$

for $\alpha = P$, F, PF; it is only necessary that T_α be a location statistic. Now it is well-known and easy to show that, for each α,

$$f_{T_\alpha|C_\alpha}(t_\alpha|c_\alpha;\theta) = \frac{\prod_1^{n_\alpha} g(y_{(\alpha j)}-\theta)}{\int \prod_1^{n_\alpha} g(y_{(\alpha j)}-t)dt}$$

and

$$f_{C_\alpha}(c_\alpha) = n_\alpha! \int \prod_1^{n_\alpha} g(y_{(\alpha j)}-t)dt \ .$$

Suppose that conditions (i) - (iii) of (3.11) hold with δ_P, δ_R and δ_{PR} null. Also suppose that $s_\omega = (T_\omega, C_\omega)$ with C_ω ancillary for ω = P, R, PR. Clearly, C_{PR} is a function of (C_P, C_R, T_P, T_R) involving an ancillary function of (T_P, T_R), and T_{PR} is a function of (T_P, T_R). The statistic C_P contains no information about θ, and hence none about s_R, so we condition on C_P and write

<u>Definition 3</u> Under the above conditions,

$$FLIK(s_R|s_P) = f_{C_R}(c_R) f_{T_P|C_{PR},C_P}(t_P|s_*(s_P,s_R),c_P) \qquad (3.16)$$

$$= f_{C_R}(c_R) FLIK_C(t_P|s_P,s_R) .$$

When C_{PR} is null, this reduces to Definition 2.

The following familiar example illustrates this definition.

<u>Example 3.5</u> Let $Y_1,\dots,Y_N$ be i.i.d. with continuous p.d.f. $g(y-\theta)$, supported on the whole real line, with $Y = (Y_1,\dots,Y_{n_P})$ and $Z = (Y_{n_P+1},\dots,Y_{n_P+n_R})$, $N = n_P+n_R$. ... the ordered values of Y, Z and (Y,Z) by $Y_{(\omega j)}$, $j=1,\dots,n_\omega$, with ω = P, R and PR respectively. Then in general $s_\omega = Y_{(\omega \cdot)}$ and we may take

$$T_\omega = n_\omega^{-1} \sum Y_{(\omega j)},\quad C_\omega = (Y_{(\omega,j+1)} - Y_{(\omega,j)};\ j=1,\dots,n_\omega-1)$$

for ω = P, R, PR; it is only necessary that T be a location statistic. Now it is well-known and easy to show that, for each ω,

$$f_{T_\omega|C_\omega}(t|c_\omega;\theta) = \frac{\prod_{j=1}^{n_\omega} g(y_{(\omega j)}-\theta)}{\int \prod_{j=1}^{n_\omega} g(y_{(\omega j)}-t)dt}$$

and

$$f_{C_\omega}(c_\omega) = n_\omega \int \prod_{j=1}^{n_\omega} g(y_{(\omega j)}-t)dt .$$

It then follows readily from Definition 3 that

$$\text{PLIK}(s_F|s_P) = f_{C_F(c_F)} f_{T_P|C_P}(t_P|c_P) f_{T_F|C_F}(t_F|c_F) \div f_{T_{PF}|C_{PF}}(t_{PF}|c_{PF})$$

$$= n_F! \frac{\int\{\prod_1^N g(y_j-u)\}du}{\int\{\prod_1^{n_P} g(y_k-v)\}dv} \quad ; \qquad (3.16)$$

for the unordered vector Z the factor $n_F!$ disappears.

Note that (3.16) is formally identical to the pivotal (hence fiducial) density. The importance of conditioning is easily understood in this case: essentially we are predicting the difference of averages $T_P - T_F$, which, given the configuration ancillaries C_P and C_F, has a known distribution with mean not necessarily equal to zero, for example. Our inference must surely use this distributional information. □□□

The full implications of the conditional predictive likelihood definition are not yet realized. However, two facts are reasonably clear, particularly in light of the last example. Firstly, unless C_{PF} is null (as in Section 3.3), the Bayes result (3.14) does not in general hold. Secondly, there is a very close relationship with pivotal inference (Section 2.2), in the sense that the conditional predictive likelihood seems to be derived solely from C_{PF}.

in which boldface variables mean matrices and

$$\mathrm{var}\{\hat{\theta}_{ML}\} = \mathbf{C}^{-1}(\theta_0)\,\mathbf{B}(\theta_0)\,\mathbf{C}^{-1}(\theta_0) \doteq \frac{1}{N}\,\mathbf{C}_{NN}^{-1}(\theta_0)\,\mathbf{B}_{NN}(\theta_0)\,\mathbf{C}_{NN}^{-1}(\theta_0) \tag{3.44}$$

$$= \mu_N^{-1}\cdot\frac{\int_{-\pi}^{\pi}\Phi^2(\lambda)\,d\lambda}{\left[\int_{-\pi}^{\pi}\Phi(\lambda)\,d\lambda\right]^2} = \mu_N^{-1}\cdot\frac{\int_{-\pi}^{\pi}\Phi^2(\lambda)\,d\lambda}{\left[\int_{-\pi}^{\pi}\Phi(\lambda)\,d\lambda\right]^2}$$

Note that (3.44) is formally identical to the hybrid (pure bandlimited) density. The equivalence to sometimes-bandlimited data is established in this case: essentially the only difference between ρ^{hy} and ρ^{sb} is that $\mu_N = \mu_N^{\mathrm{hy}}$ depends on the constellation probabilities p_k and q_k; but $\mu_N^{\mathrm{sb}} = \mu_N^{\mathrm{hy}}$ depends only on the constellation probabilities p_k. Hence, for example, the sometimes-bandlimited data can be viewed as pure bandlimited data with fewer effective symbols. □□□

The last two equations are to be considered heuristic. However, they are mathematically correct, provided the limit can be taken. Moreover, the above expressions ρ^{sb} assume, without loss of generality, that $\Phi(\lambda)$ does not vanish identically on any interval (see also section 3.6); and pure hybrid data (section 3.5.2) does not in general satisfy this condition. Consequently, under the above assumptions, ρ^{hy} is strictly greater than ρ^{sb} for all N (see section 3.5.3), and pure hybrid data is strictly more efficient than pure sometimes-bandlimited data in this sense.

4. An Empirical Analog of Predictive Likelihood

Thus far the discussion has dealt entirely with prediction in a parametric framework, the connection between data and predictand being the unknown constant θ. This classical approach yields smooth and precise inferential statements by concentrating sample information on the few defining parameters of a known distribution family. A possible disadvantage of the approach is lack of flexibility in model fitting or, more accurately, failure to mirror the pragmatic approaches of applied statistics. Some concern on this point has recently stimulated work on the cross-validatory approach to prediction and model-fitting, notably by Stone [18] and Geisser [12]. The principal idea is to compare a (point) predictor estimated from a data subset with the predictands in the remaining data. Much the same idea underlies our definition of predictive likelihood in Section 3.2, which suggests the possibility of deriving an empirical analog of predictive likelihood by data sub-sampling.

We consider here only the simple case of one-dimensional exchangeable variables and prediction of a single outcome. Suppose that $Y_1,\ldots,Y_{n+1}$ are exchangeable, that we observe $Y_1=y_1,\ldots,Y_n=y_n$ and wish to predict Y_{n+1}. Now let the statistic $s(y_1,\ldots,y_n)$ be chosen as a single data summary, presumably subjectively, so that differences in s will reflect important changes in the data. In the parametric framework $s(\cdot)$ would be dictated by the minimal sufficiency requirement, but that is not so here; indeed one might well choose two or more statistics $s_1,s_2,\ldots$. Because of the assumed exchangeability of $Y_1,\ldots,Y_n$, $s(\cdot)$ is taken to be permutation-invariant. For this situation it is easy to develop an empirical analog of (3.3), as follows.

Boolean Algebra Extension to Normal Logic

A multivariate probability model can be developed that is based on the concept of a Boolean algebra. The Boolean algebra is a mathematical structure that consists of a set of elements and two binary operations, usually denoted by $\cdot$ and $+$, that satisfy certain axioms. The elements of the set can be thought of as propositions, and the operations $\cdot$ and $+$ can be thought of as logical AND and OR, respectively. The axioms of the Boolean algebra include the commutative, associative, and distributive laws, as well as the existence of identity elements and complements.

In the context of probability theory, the Boolean algebra can be used to represent the relationships between events. The elements of the Boolean algebra can be thought of as events, and the operations $\cdot$ and $+$ can be thought of as the intersection and union of events, respectively. The complement of an event is the event that does not occur.

The probability of an event can be represented by a function that maps the elements of the Boolean algebra to the interval $[0,1]$. This function must satisfy certain axioms, such as the probability of the entire sample space being 1, and the probability of the union of mutually exclusive events being the sum of their probabilities.

The Boolean algebra extension to normal logic allows for the representation of uncertainty in logical statements. In normal logic, statements are either true or false, but in the Boolean algebra extension, statements can have a probability between 0 and 1. This allows for the representation of uncertainty in logical reasoning.

The Boolean algebra extension to normal logic can be used in a variety of applications, such as artificial intelligence, machine learning, and decision making under uncertainty. It provides a powerful framework for representing and reasoning about uncertainty in logical statements.

One of the key advantages of the Boolean algebra extension to normal logic is its ability to handle incomplete information. In many real-world situations, we do not have complete information about the world, and we must make decisions based on incomplete information. The Boolean algebra extension allows us to represent this incomplete information and to reason about it in a principled way.

Another advantage of the Boolean algebra extension to normal logic is its ability to handle inconsistent information. In many real-world situations, we may have conflicting information about the world, and we must make decisions based on this conflicting information. The Boolean algebra extension allows us to represent this conflicting information and to reason about it in a principled way.

The Boolean algebra extension to normal logic is a powerful framework for representing and reasoning about uncertainty in logical statements. It has a wide range of applications, and it is a valuable tool for anyone who needs to make decisions under uncertainty.

Let $D_j = \{y_k : k=1,\ldots,n, k\neq j\}$, $j=1,\ldots,n$ and $D_{jk} = \{y_i : i=1,\ldots,n, i\neq j, i\neq k\}, k\neq j$. Then by assumption all n subsamples D_j are equally likely, and all (n-1) subsamples D_{jk} within D_j are equally likely. We now have, in obvious notation, n(n-1) pairs $(s(D_{jk}),s(D_j))$ which correspond to (S_P,S_{PF}) as defined in Section 3.3. Suppose that there are r_n distinct values $t_1,\ldots,t_{r_n}$ of $s(D_{jk})$, c_n distinct values $u_1,\ldots,u_{c_n}$ of $s(D_j)$, and define the r_n x c_n matrix $\underset{\sim}{m}$ to have (a,b) element equal to the frequency of (t_a,u_b) in the totality of $\{(s(D_{jk}),s(D_j)), 1 \leq k \neq j \leq n\}$ of nested subsamples. Then the matrix $\underset{\sim}{m}$ gives an empirical analog of (3.3), namely the subsampling frequency

$$\text{FREQ}(S_P=t_a \,|\, S_{PF}=u_b; n_P=n-2, n_F=1) = \frac{m_{ab}}{m_{+b}} , \tag{3.17}$$

where $m_{+b} = \sum_1^{r_n} m_{ab}$. Now since u_b and t_a determine the (n-1) st. value of y_j, say $y^*_{ab} = y^*(t_a,u_b)$, the right hand side of (3.17) may be denoted $\tilde{m}(t_a,y^*_{ab};n-2)$. By extrapolation we have

Definition 4 If (3.17) is equal to $\tilde{m}(t_a,y^*_{ab};n-2)$, then the empirical predictive likelihood of the event $Y_{n+1} = y^*$, given that $s(y_1,\ldots,y_n) = s$, is

$$\text{EPLIK}(y^* \,|\, s) = \tilde{m}(s,y^*;n) , \tag{3.18}$$

with support equal to $\{y^*_{ab} : a=1,\ldots,r_n; b=1,\ldots,c_n\}$.

Example 4.1 Let the data consist of exchangeable Bernoulli trials, each y_j being 0 or 1, and suppose that $s(y_1,\ldots,y_n) = \Sigma y_j = s$. A small calculation shows that $r_n = 3$, $c_n = 2$, $(t_1,t_2,t_3) = (s-2,s-1,s)$, $(u_1,u_2) = (s-1,s)$ and

$$\underset{\sim}{m} = \begin{pmatrix} s(s-1) & 0 \\ s(n-s) & s(n-s) \\ 0 & (n-s)(n-s-1) \end{pmatrix}.$$

Let $D_j = \{y_i : i=1,\ldots,n,\ i\neq j\}$, $j=1,\ldots,n$ and $D_{jk} = \{y_i : i=1,\ldots,n,\ i\neq j,\ i\neq k\}$, $k\neq j$. Then by assumption all n subsamples D_j are equally likely, and all $(n-1)$ subsamples D_{jk} within D_j are equally likely. We now have, in obvious notation, $n(n-1)$ pairs $(s(D_{jk}), s(D_j))$ which correspond to (S_{nK}, S_n) as defined in Section 3.3. Suppose that there are r_n distinct values $t_1,\ldots,t_{r_n}$ of $s(D_{jk})$, c_n distinct values $u_1,\ldots,u_{c_n}$ of $s(D_j)$, and define the $r_n \times c_n$ matrix M to have (a,b) element equal to the frequency of (t_a,u_b) in the totality of $\{(s(D_{jk}), s(D_j)),\ 1 \le j \neq k \le n\}$ of nested subsamples. Then the matrix M gives an empirical analog of (3.3), namely the subsampling frequency

$$\mathrm{EMPR}(S_{nK}=t_a \mid S_n=u_b;\ n-1, n-2) = \frac{m_{ab}}{m_{\cdot b}}, \qquad (3.16)$$

where $m_{\cdot b} = \sum_a m_{ab}$. Now since u_b and t_a determine the $(n-1)$st value of y_i, say $y^*_{ab} = y^*(t_a,u_b)$, the right hand side of (3.16) may be denoted $\tilde{m}(t_a, y^*_{ab}; n-2)$. By extrapolation we have

<u>Definition 4</u> If (3.16) is equal to $\tilde{m}(t_a, y^*_{ab}; n-2)$, then the empirical predictive likelihood of the event $Y_{n+1} = y^*$, given that $s(y_1,\ldots,y_n) = s$, is

$$\mathrm{EMPR}(y^* \mid s) = \tilde{m}(s, y^*; n), \qquad (3.17)$$

with support equal to $\{y^*_{ab} : a=1,\ldots,r_n;\ b=1,\ldots,c_n\}$.

<u>Example 4.1</u> Let the data consist of exchangeable Bernoulli trials, each y_i being 0 or 1, and suppose that $s(y_1,\ldots,y_n) = \sum y_i = s$. A small calculation shows that $r_n = 3$, $c_n = 2$, $(t_1,t_2,t_3) = (s-2, s-1, s)$, $(u_1,u_2) = (s-1,s)$ and

$$M = \begin{pmatrix} s(s-1) & 0 \\ s(n-s) & s(n-s) \\ 0 & (n-s)(n-s-1) \end{pmatrix}.$$

Then (3.17) and (3.18) give

$$\text{EPLIK}(y^*|s) = \left(\frac{n-s+1}{n+1}\right)^{1-y^*}\left(\frac{s+1}{n+1}\right)^{y^*} \quad , \; y^* = 0,1 \quad ,$$

which is exactly the parametric result (3.4) with $n_F = 1$. This is not very surprising, since our assumptions almost imply the binomial model. □□□

Sampling two deep is necessary because in the definition (3.3) it is S_{PF} that varies. For general $n_F \geq 1$ the subsampling is of subsets size $n_P - 2n_F$ from subsets size $n_P - n_F$. The assumption of a particular form for $s(\cdot)$ is very close to a parametric assumption in Example 4.1, but not in general; for example Σy_j is sufficient for several continuous families.

Example 4.2 Suppose that $y_1,\ldots,y_n$ are distinct values, and let $s(y_1,\ldots,y_n) = y_{(n)}$, the largest order statistic, which is the lowest known upper bound on observations. Then $r_n = 3$, $c_n = 2$, $(t_1,t_2,t_3) = (y_{(n-2)},y_{(n-1)},y_{(n)})$, $(u_1,u_2) = (y_{(n-1)},y_{(n)})$ and

$$\underset{\sim}{m} = \begin{pmatrix} 1 & 0 \\ n-2 & n-1 \\ 0 & (n-1)(n-2) \end{pmatrix}$$

For prediction of one further outcome y^*, (3.17) and (3.18) give

$$\text{EPLIK}(y^*|y_{(n)}) = \begin{cases} 1 - (n+1)^{-1}, & y^* \leq y_{(n)} \\ (n+1)^{-1} \quad , & y^* > y_{(n)} \end{cases} .$$

This result is to be compared with the result for the uniform case in Example 3.3: here the likelihood beyond $y_{(n)}$ is not distributed, since no density is assumed. □□□

Many interesting situations, such as that involving $s = \Sigma y_j$, encounter the difficulty that subsampled values of s may be distinct, leading to a matrix $\underset{\sim}{m}$ consisting of zeros and ones. If the empirical likelihood were to be generally useful, some smoothing device (acting on y or $\underset{\sim}{m}$) would be required. Our interest here is solely in noting the possibility of the cross-validation analog of Definition 1.

5. Concluding Remarks

In Section 2 some negative aspects of conventional sampling-theory prediction regions are emphasized, in order to give point to the subsequent discussion of likelihood. Clearly the sampling-theory methods are useful, and sensible confidence regions have usually been proposed that would agree with ordering by predictive likelihood PLIK. Some interesting work on prediction confidence regions may be found in [19], [6], [2], [13]; the latter is an exposition of invariance theory in pivotal prediction.

The treatment of predictive likelihood in Sections 3.3-4 clearly ignores some very interesting technical problems connected with the structure of sufficient statistics. There is currently much interest in such problems, particularly among Scandinavian statisticians; see particularly Lauritzen's work in [14], [15].

6. Acknowledgements

I particularly thank Steffen Lauritzen and Stephen Stigler for critical comments on an earlier draft of this article; Lauritzen suggested that the result (3.8) is true. Much of the work reported here was supported by NSF Grant No. NSF-MPS75-08778.

5. Concluding Remarks

In Section 2 some negative aspects of conventional sampling-theory prediction regions are emphasised, in order to give point to the subsequent discussion of likelihood. Clearly the sampling-theory methods are useful, and sensible confidence regions have usually been proposed that would agree with ordering by predictive likelihood PLN. Some interesting work on prediction confidence regions may be found in [19], [6], [2], [13]; the latter is an exposition of invariance theory in physical prediction.

The treatment of predictive likelihood in Sections 3.3-4 clearly ignores some very interesting technical problems connected with the structure of sufficient statistics. There is currently much interest in such problems, particularly among Scandinavian statisticians; see particularly Lauritzen's work in [14], [15].

6. Acknowledgements

I particularly thank Steffen Lauritzen and Stephen Stigler for critical comments on an earlier draft of this article. My thoughts about prediction were prompted by the material in Section 7.5 of [7].
Much of the work reported here was supported by NSF Grant No.

TABLE 1

Prior measures for which PLIK $\propto$ Bayes posterior density

Model	Parameter θ	Prior Measure
i.i.d. MN(μ, Σ) in q dimensions	μ, Σ^{-1}	$\lvert\Sigma\rvert^{\frac{1}{2}q} d\mu\, d\Sigma^{-1}$
same, q=1, $\Sigma=\sigma^2$	μ, σ^2	$\sigma d\mu\, d\sigma$
i.i.d. Bernoulli with pr(Y=1)=θ=pr(Y=0)		
(a) fixed number of trials	θ	$d\theta$
(b) fixed number of ones	θ	$\theta^{-2} d\theta$
Poisson process with constant event rate λ		
(a) i.i.d. counts	λ	$d\lambda$
(b) i.i.d. inter-event times	λ	$\lambda^{-2} d\lambda$
i.i.d. Uniform on [0, θ]	θ	$\theta^{-1} d\theta$

REFERENCES

[1] Barnard, G.A. (1975). New foundations of statistical inference. Unpublished lecture notes, University of Minnesota.

[2] Cox, D.R. (1974). Prediction intervals and empirical Bayes confidence intervals. (Unpublished).

[3] Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman - Hall.

[4] Edwards, A.W.F. (1974). A problem in the doctrine of chances. Proc. Conference on Foundational Questions in Statistical Inference, eds. Barndorff-Nielsen, O. et al. University of Aarhus.

[5] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) Ann. Stat. 3, (to appear).

[6] Faulkenberry, G.D. (1973). A method of obtaining prediction intervals. J. Amer. Statist. Assoc. 68, 433-435.

[7] de Finetti, B. (1937). La Prévision: ses lois logiques, ses sources subjectives. Ann. Inst. H. Poincaré, 7, 1-68.

[8] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. Phil. Trans. Roy. Soc. London Ser. A 222, 309-368.

[9] Fisher, R.A. (1935). The fiducial argument in statistical inference. Ann. Eug. 6, 391-398.

[10] Fisher, R.A. (1973). Statistical Methods and Scientific Inference. (3rd Ed.) New York: Hafner.

[11] Geisser, S. (1971). The inferential use of predictive distributions. In Foundations of Statistical Inference (V.P. Godambe and D.A. Sprott, eds.) pp. 456-469. Toronto, Montreal: Holt, Rinehart and Winston.

[12] Geisser, S. (1975). The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 79, (to appear).

[13] Hora, R.B. and Buehler, R.J. (1967). Fiducial theory and invariant Prediction. Ann. Math. Statist. 38, 795-801.

[14] Lauritzen, S.L. (1974). Sufficiency, prediction and extreme models. Scand. J. Statist. 1, 128-134.

[15] Lauritzen, S.L. (1975). General exponential models for discrete observations. Scand. J. Statist. 2, 23-33.

[1] Ломакина, Л.А. (1962). Тем хорошо знать то, что хорошо забыто. Литературное
путешествие-последнее время, Литературное.

[2] Кон, Н.Н. (1964). Взаимодействие литературы без вмешательства редактора.
Москва: Наука - Дальний.

[3] Кон, Н.Н. (1964). Публикация литературных. Москва: Наука.

[4] Кравчук, Р.И. (1964). К вопросу об отношении между литературой
и обществом. Современность и литературный процесс. Литературное
творчество. Вып. 4. с. 45-57. Ленинград: Наука.

[5] Кон, Н.Н. (1964). Взаимодействие литературы без вмешательства редактора.
Москва: Наука - Дальний.

[6] Кравчук, Р.И. (1964). К вопросу об отношении между литературой
и обществом. Современность и литературный процесс. Литературное
творчество. Вып. 4. с. 45-57. Ленинград: Наука.

[7] Кон, Н.Н. (1964). Взаимодействие литературы без вмешательства редактора.
Москва: Наука - Дальний.

[8] Кравчук, Р.И. (1964). К вопросу об отношении между литературой
и обществом. Современность и литературный процесс. Литературное
творчество. Вып. 4. с. 45-57. Ленинград: Наука.

[9] Кон, Н.Н. (1964). Взаимодействие литературы без вмешательства редактора.
Москва: Наука - Дальний.

[10] Кравчук, Р.И. (1964). К вопросу об отношении между литературой
и обществом. Современность и литературный процесс. Литературное
творчество. Вып. 4. с. 45-57. Ленинград: Наука.

[11] Кон, Н.Н. (1964). Взаимодействие литературы без вмешательства редактора.
Москва: Наука - Дальний.

[12] Кравчук, Р.И. (1964). К вопросу об отношении между литературой
и обществом. Современность и литературный процесс. Литературное
творчество. Вып. 4. с. 45-57. Ленинград: Наука.

[16] Pearson, K. (1920). The fundamental problem of practical statistics. Biometrika 13, 1-16.

[17] Pearson, K. (1921). Note on the 'fundamental problem of practical statistics'. Biometrika 16, 190-193.

[18] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, (with discussion). J. Roy. Statist. Soc. B 36, 111-147.

[19] Weiss, L. (1955). A note on confidence sets for random variables. Ann. Math. Statist. 26, 142-144.

[20] Whittle, P. (1955). Some distribution and moment formulae for the Markov chain. J. Roy. Statist. Soc. B 17, 235-242.

Appendix: Consistency of PLIK

In Section 3.2 we refer without proof to two general consistency properties of PLIK in the i.i.d. exponential family case. A stronger result is stated without proof in (3.8). The lemma below and its corollaries give proofs of those results. For simplicity θ is taken to be one-dimensional.

Lemma Let $Y_1,\ldots,Y_{n_{PF}}$ be i.i.d. with p.d.f.

$$f_Y(y;\theta) = \exp\{-\theta y + c(\theta) + d(y)\} ,$$

$S_P = \sum_1^{n_P} Y_j$, $S_F = \sum_{n_P+1}^{N_{PF}} Y$, $S_{PF} = S_P + S_F$, $n_{PF} = n_P + n_F$. Then with n_F fixed, $n_P \to \infty$,

$$f_{S_P|S_{PF}}(t|s+t) = f_{S_F|S_{PF}}(s|s+t) = f_{S_F}(t;\hat{\theta}_P) + O_p(n_P^{-1}), \quad (A.1)$$

where $\hat{\theta}_\alpha$ is the maximum likelihood estimate determined by $S_\alpha (\alpha = P,F,PF)$.

Corollary Under the conditions of the Lemma, with PLIK given by Definition 1,

(a) as $n_P \to \infty$ $\mathrm{PLIK}(S_F|S_P) = f_{S_F}(s_F;\theta) + O_p(n_P^{-\frac{1}{2}})$

(b) as $n_F \to \infty$ $\mathrm{PLIK}(S_F|S_P) = f_{S_P}(s_P;\theta) + O_p(n_F^{-\frac{1}{2}})$

Proof of Lemma The p.d.f. of S_α may be denoted

$$g_{n_\alpha}(s_\alpha;\theta) = D(s_\alpha,n_\alpha) \exp\{-\theta s_\alpha + n_\alpha c(\theta)\} \quad (\alpha=P,F,PF).$$

Then (A.1) states that

$$f_{S_P|S_P+S_F}(s_P|s_P+t) = D(t,n_F) \exp\{-\hat{\theta}_P t + n_F c(\hat{\theta}_P)\} , \quad (A.2)$$

where $$S_\alpha + n_\alpha c'(\hat{\theta}_\alpha) = 0 \qquad (A.3)$$

uniquely defines $\hat{\theta}_\alpha$. By definition

$$f_{S_P|S_P+S_F}(s|s+t) = f_{S_F|S_P+S_F}(t|s+t) = \frac{D(s,n_P)D(t,n_F)}{D(s+t,n_P+n_F)} \quad ,$$

which can be re-expressed as

$$f_{S_P|S_P+S_F}(s|s+t) = D(t,n_F)\frac{g_{n_P}(s;\hat{\theta}_P)}{g_{n_{PF}}(s+t;\hat{\theta}_{PF})}\exp\{-n_P c(\hat{\theta}_P)+n_{PF}c(\hat{\theta}_{PF})+\hat{\theta}_P S-\hat{\theta}_{PF}(s+t)\}. \qquad (A.4)$$

But from (A.3) we deduce, by expansion, that

$$\hat{\theta}_P - \hat{\theta}_{PF} = \frac{n_P(s+t) - n_{PF}S}{n_P n_{PF}\{c''(\theta)+0_p(n_P^{-\frac{1}{2}})\}} = 0_p(n_P^{-1}) \quad , \qquad (A.5)$$

so that (A.4) becomes

$$f_{S_P|S_P+S_F}(s|s+t) = g_{n_F}(t;\hat{\theta}_P)\{1+0_p(n^{-1})\}\frac{g_{n_P}(s;\hat{\theta}_P)}{g_{n_{PF}}\{s+0(1);\hat{\theta}_P+0_p(n_P^{-1})\}} \, .$$

It therefore remains to show that

$$\frac{g_{n_P}(s;\hat{\theta}_P)}{g_{n_{PF}}\{s+0(1);\hat{\theta}_P+0_p(n_P^{-1})\}} = 1 + 0_p(n_P^{-1}) \, .$$

This follows by applying (i) $f_{X+0_p(n^{-1})}(x) = f_X(x) + 0(n^{-1})$;

(ii) $g_n(s+0(1);\theta) = g_n(s;\theta) + 0(n^{-1})$, (iii) $g_n(s;\theta+0(n^{-1})) = g_n(s;\theta) + 0(n^{-1})$.

Proof of Corollary The first result (a) follows directly from (A.1) and (iii) above because $\hat{\theta}_P = \theta + 0_p(n_P^{-\frac{1}{2}})$. The second result (b) is a consequence of (A.1) with the roles of (S_P,n_P) and (S_F,n_F) interchanged.

The above arguments extend directly to the finite-dimensional parameter case, and to the exponential family linear model case. It is conjectured that corresponding results hold for Definitions 2 and 3, but this has not been proved.

Then (3.17) and (3.18) give

$$\text{EPLIK}(y^*|s) = \left(\frac{n-s+1}{n+1}\right)^{1-y^*}\left(\frac{s+1}{n+1}\right)^{y^*} \quad , \; y^* = 0,1 \quad ,$$

which is exactly the parametric result (3.4) with $n_F = 1$. This is not very surprising, since our assumptions almost imply the binomial model. □□□

Sampling two deep is necessary because in the definition (3.3) it is S_{PF} that varies. For general $n_F \geq 1$ the subsampling is of subsets size $n_P - 2n_F$ from subsets size $n_P - n_F$. The assumption of a particular form for $s(\cdot)$ is very close to a parametric assumption in Example 4.1, but not in general; for example Σy_j is sufficient for several continuous families.

Example 4.2 Suppose that $y_1,\ldots,y_n$ are distinct values, and let $s(y_1,\ldots,y_n) = y_{(n)}$, the largest order statistic, which is the lowest known upper bound on observations. Then $r_n = 3$, $c_n = 2$, $(t_1,t_2,t_3) = (y_{(n-2)},y_{(n-1)},y_{(n)})$, $(u_1,u_2) = (y_{(n-1)},y_{(n)})$ and

$$\underset{\sim}{m} = \begin{pmatrix} 1 & 0 \\ n-2 & n-1 \\ 0 & (n-1)(n-2) \end{pmatrix}$$

For prediction of one further outcome y^*, (3.17) and (3.18) give

$$\text{EPLIK}(y^*|y_{(n)}) = \begin{cases} 1 - (n+1)^{-1}, & y^* \leq y_{(n)} \\ (n+1)^{-1}, & y^* > y_{(n)} \end{cases} .$$

This result is to be compared with the result for the uniform case in Example 3.3: here the likelihood beyond $y_{(n)}$ is not distributed, since no density is assumed. □□□

and (11.6) and (8.6) give

$$\max(\mathbf{e}_{n-1}|\theta) = \binom{n-1}{L-1}\gamma^{L-1}\binom{n-L}{L-1}\beta^{2L-2}, \quad L \geq 1, \; n \geq 0$$

from which we find $L = \mu_E$ after differentiating and equating to zero. Hence the Lagrange multiplier method is equivalent to the maximum likelihood method in this case.

(6.6) established that the maximum likelihood estimate of θ is the reciprocal of the sample mean. Hence the Lagrange multiplier estimate of θ is also the reciprocal of the sample mean. Thus the Lagrange multiplier method is equivalent to the maximum likelihood method in this case.

Let us now consider the problem of estimating the parameters $\gamma_1, \ldots, \gamma_L$ of the Markov chain $\{\mathbf{e}_n\}$ when L is known. Let $\mathbf{e}^{(m)} = (e_m, \ldots, e_{m+L-1})$ denote the mth state of the chain. Then $\mathbf{e}^{(m)} = (e_m, e_{m+1}, \ldots, e_{m+L-1})$ can be regarded as a Markov chain with L possible states $\{\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(L)}\}$ and transition probabilities

$$\Pr(\mathbf{e}^{(m)} \to \mathbf{e}^{(n)}) = \gamma_n^{(L-n)}\gamma_{n-1}^{(L-n+1)}\cdots\gamma_1^{(L-1)}(1-\gamma_n)^{(n)}(1-\gamma_{n-1})^{(L-n+1)}\cdots(1-\gamma_1)^{(L-1)}$$

$$= \begin{pmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & \ddots & \ddots \\ 1-\gamma_L & & & \gamma_L \\ \vdots & & & \vdots \\ 1-\gamma_2 & & & \gamma_2 \\ 1-\gamma_1 & & & \gamma_1 \end{pmatrix} = \mathbf{M}$$

say. Then (11.6) and (8.6) give

$$\max(\mathbf{e}_{n-1}|\theta) = \binom{n-1}{L-1}\gamma^{L-1}\binom{n-L}{L-1}\beta^{2L-2}, \quad L \geq 1, \; n \geq 0$$

from which we find $L = \mu_E$ after differentiating and equating to zero. Hence the Lagrange multiplier method is equivalent to the maximum likelihood method in this case.

(6.6) established that the maximum likelihood estimate of θ is the reciprocal of the sample mean. Hence the Lagrange multiplier estimate of θ is also the reciprocal of the sample mean. Thus the Lagrange multiplier method is equivalent to the maximum likelihood method in this case.

Let us now consider the problem of estimating the parameters $\gamma_1, \ldots, \gamma_L$ of the Markov chain $\{\mathbf{e}_n\}$ when L is known. Let $\mathbf{e}^{(m)} = (e_m, \ldots, e_{m+L-1})$ denote the mth state of the chain. Then $\mathbf{e}^{(m)} = (e_m, e_{m+1}, \ldots, e_{m+L-1})$ can be regarded as a Markov chain with L possible states $\{\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(L)}\}$ and transition probabilities

$$\Pr(\mathbf{e}^{(m)} \to \mathbf{e}^{(n)}) = \gamma_n^{(L-n)}\gamma_{n-1}^{(L-n+1)}\cdots\gamma_1^{(L-1)}(1-\gamma_n)^{(n)}(1-\gamma_{n-1})^{(L-n+1)}\cdots(1-\gamma_1)^{(L-1)}$$

$$= \begin{pmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & \ddots & \ddots \\ 1-\gamma_L & & & \gamma_L \\ \vdots & & & \vdots \\ 1-\gamma_2 & & & \gamma_2 \\ 1-\gamma_1 & & & \gamma_1 \end{pmatrix} = \mathbf{M}$$

say. Then (11.6) and (8.6) give

Many interesting situations, such as that involving $s = \Sigma y_j$, encounter the difficulty that subsampled values of s may be distinct, leading to a matrix $\underset{\sim}{m}$ consisting of zeros and ones. If the empirical likelihood were to be generally useful, some smoothing device (acting on y or $\underset{\sim}{m}$) would be required. Our interest here is solely in noting the possibility of the cross-validation analog of Definition 1.